United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

RD Research Report
Number RD-99-04

August 1999

A Methodology for Small Area Estimation with Special Reference to a One-Number Agricultural Census and Confidentiality: Results for Selected Major Crops and States

Daniel A. Griffith

A METHODOLOGY FOR SMALL AREA ESTIMATION, WITH SPECIAL REFERENCE TO A ONE-NUMBER AGRICULTURAL CENSUS AND CONFIDENTIALITY: RESULTS FOR SELECTED MAJOR CROPS AND STATES, by Daniel A. Griffith, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC, August 1999, NASS Research Report RD-99-04.

ABSTRACT

This report describes a flexible methodology for calculating spatial autoregressive model-based small geographic estimates, illustrating its utility with substitution computations for the suppression code (D) in order to allow more complete data tabulations to be released. Post-stratification figures reported here differ from those published in the 1997 Census of Agriculture, which was conducted by the National Agricultural Statistics Service (NASS). Estimated census totals differ somewhat, for several agricultural items of interest, from official estimates based upon the ongoing NASS survey program and administrative data. These discrepancies can be attributed to nonsampling errors, timing and process differences. One such source of nonsampling error is underenumeration, resulting from some small farms not being on the census mail list. Figures reported here have been adjusted for such errors. Once the best estimates of agricultural commodities have been established, another concern is preserving confidentiality while releasing as much of the data as possible. This issue arises when these data are post-stratified into counties, for small geographic area tabulations. The text describes applications of the estimation methodology to principal agricultural commodities for Michigan and Tennessee that have been adjusted in order to align the agricultural census results with reliable official totals. It also describes comparisons between the spatial autoregressive and other model-based estimates. One important finding is that currently anyone with a working knowledge of estimators like those discussed here already is capable of generating quite accurate synthetic values from public domain databases, and then substituting them for suppressed values.

KEY WORDS

CALJACK; confidentiality; EM algorithm; imputation; one-number census; small area estimation; spatial autoregressive model; suppression.

ACKNOWLEDGMENTS

The research whose findings are summarized in this report was made possible by an ASA/USDA-NASS fellowship during January-August, 1999. The author thanks Ron Bosecker for his leadership in this effort and for constructive comments on earlier drafts of this report. In addition, Chadd Crouse is gratefully acknowledged for furnishing a description of and results from CALJACK.

TABLE OF CONTENTS

SUMMARY

The National Agricultural Statistics Service (NASS) now conducts the periodic national census of agriculture. Data collected are from individual farms and must be held in confidence. Accordingly, any post-stratifications of these data that might breach this confidentiality are not released. Suppression of data that are post-stratified by county create maps with holes in them. In addition, to properly account for a given post-stratified agricultural commodity aggregate, such as county or state totals, NASS adjusts the weighting of the individual farm data to compensate for nonresponse. Weights currently being explored that are generated with a SAS macro known as CALJACK compensate for farms not on the list sampling frame, list duplication and misclassification of farm status, and can further realign the individual farm data with known "official" state or agricultural district totals, if desired.

Provisions of Title 13 of the U.S. code maintain that no data are to be published that would disclose the operations of an individual farm. But the number of farms in a county post-stratification is not considered a release of confidential information, and is published. A first-step in model-based estimation of suppressed county values, then, is to exploit covariation between the distributions of an agricultural commodity and its accompanying number of farms across counties. A second covariate is the presence or absence of production in a county. A simple binary indicator variable can distinguish between the two categories of this situation. Thus, either a bivariate or a trivariate regression model can be used to estimate suppressed values. The estimation procedure is EM (i.e., estimation-maximization) in nature: suppressed values are first estimated using the regression equation obtained with the unsuppressed values; then these predicted suppressed values (i.e., the estimates) are treated as observed values, and the regression parameters re-estimated (maximum likelihood estimation). The repeated substitution of updated predicted values (the E part of the algorithm) followed by re-estimation of the regression equation (the M part of the algorithm) continues until the parameter estimates stabilize.

Methodology presented in this report uses information contained in NASS "official" totals to modify this estimation in such a way that the proportion of a given total is estimated for each county having a suppressed value. Now the regression must be nonlinear, and each suppressed value must be constrained by some minimum. Because only positive values of production are suppressed in publication, the minimum used for this report is 1. Of course, any feasible minimum value could be set. An additional source of information for exploitation is spatial autocorrelation latent in the county geographic distribution of an agricultural commodity. This correlation arises because nearby farms tend to produce similar agricultural commodities.

Four model-based estimators are evaluated in this report. One is the unsophisticated weighted average, with weights being the relative numbers of farms. A second is the constrained EM-type regression estimator, which focuses on the bi- or tri-variate relationship between an agricultural commodity and the number of farms producing it. A third is the pure spatial autoregressive estimator built upon spatial autocorrelation, employing the simultaneous autoregressive (SAR) model and the simple binary adjacency geographic structure. And a fourth is the hybrid constrained

EM-type SAR estimator.

Because of their availability, major crops for Michigan and Tennessee have been used for empirical evaluation purposes.

The methodology presented in this report is sufficiently flexible that it can be used with census results whether or not they have been realigned with official estimates through CALJACK calibration.

Major conclusions stemming from this exercise include:

(1) useful small geographic area estimates are obtainable that can be devised to preserve confidentiality, allow model-based imputations to be reported, and should prove informative to the NASS debate concerning a "one-number" census utilizing numerical calibrations like CALJACK adjustments;

(2) the methodology outlined in this report is sufficiently flexible that it can be employed to solve a wide range of geographic imputation problems;

(3) NASS needs to recognize that currently a knowledgeable developer and user of estimators like those discussed in this report already is capable of generating quite accurate synthetic values from public domain databases, and then substituting them for Agency-suppressed values; and,

(4) a number of instructive research projects should be undertaken in order to attain a much better understanding of the utility and limitations of model-based small geographic area estimators.

Secondary conclusions are based upon empirical findings, and include: (1) the spatial autoregressive estimator appears to perform best for major crops in Michigan and Tennessee; and, (2) one of the goals of model-based small geographic area estimation should be to preserve a certain degree of noise in generated synthetic values.

INTRODUCTION

Citro (1998) contends that "the major advance of the future that will enable the federal statistical system to remain relevant to user needs is model-based estimation for small geographic areas." The objective of this paper is to outline and evaluate a methodology for implementing this type of small area estimation for agricultural statistics, illustrating this methodology using the 1997 Census of Agriculture county-level data collected about major commodities for Michigan and Tennessee. Subsequent reports will deal with secondary and minor commodities, and with geo-demographic small area estimation problems relevant to agriculture. Because presently the list frame is not geocoded, which is a requirement for much small geographic area estimation work, the particular demonstration explored here is imputation of values that will preserve the notion of a "one number" census, maintain confidentiality of the data, and remove the (D) entries in released tabulations. The motivation for doing this is twofold: (1) to allow more complete data tabulations to be released; and, (2) rather than remain unpublished, replace suppressed values with imputed ones. Of note is that Pannekoek and de Waal (1998) also utilize the small area estimation approach for statistical disclosure control, considering the case of regional indicators being present, but ignoring the presence of latent spatial correlation.

AN APPROACH TO THE ONE-NUMBER CENSUS PROBLEM

The issue of how to handle, and hopefully reduce, differences between totals estimated from census returns and official totals estimated with the ongoing NASS survey, for the 2002 Census of Agriculture, will be discussed at great length within NASS over the next few years. To some extent these differences result from definitional and process variations that in part reflect different goals of the two efforts. The primary goal of the agricultural census is to provide very detailed information, including demographics, on agriculture at many levels of aggregation. The primary goal of ongoing NASS surveys is to forecast and estimate U.S. agricultural supply, incorporating administrative check data, such as cotton ginnings, soybean crushings and livestock slaughter data, when doing so. As much as possible, definitions will be standardized prior to the 2002 Census. After all possible changes are made to increase comparability, however, some differences will remain.

Calibration is needed to project state-level census totals adjusted for nonresponse, duplication and misclassification to finer levels of aggregation. Calibration also might be used to adjust for undercoverage. The development in this paper is couched in these latter terms, and describes the most general use of procedures that exploit all information available to the Agency (i.e., census results, estimated census adjustments, survey results and official estimates) in order to produce a single best estimate for each agricultural commodity. If the final decision is to use these procedures only to distribute census adjustments to lower levels of aggregation, though, lessons learned in optimally using these procedures should still be enlightening for that type of calibration as well.

Adjustments in collected agricultural statistics must correct for nonresponse to surveys, undercoverage of the mail list, duplication, misclassification, and other nonsampling errors. In addition, the basic calibration prob-

lem is to adjust survey[1] and census results so that they are consistent with official statistics based on externally obtained check data from various sources that give "known" commodity production totals[2]. These adjustments and this calibration need to result in a single number for small areas, whether they are geographic or non-geographic in nature. The SAS macro CALJACK[3] has been used with the data analyzed here in order to generate results that are calibrated to official statistics totals published by the National Agricultural Statistics Service (NASS), which are considered reliable for either districts in a state or the state itself. CALJACK is an upgraded version of CALMAR, which adjusts a sample by weighing individual returns such that they are calibrated to some known population. The new version of the routine also has been modified to allow calibrated totals to deviate from official totals by as much as ±5%, rather than perhaps unsuccessfully forcing exact equality between these pairs of totals as CALMAR does. This modification increases the likelihood of calibration convergence in problematic cases. In addition, each CALJACK weight is bounded, not being permitted to be less than ¼ of, or more than 4 times, its corresponding original weight. The final solution is the one closest to a set of prespecified weights, all of which might be 1.

The first step in this calibration procedure is to remove any large "outliers" from the individual farms survey data set (this is equivalent to setting the respective weights to 1). For the 1997 agricultural census data, a farm was identified as being an outlier if it either accounted for a large percentage of one or more commodities in its county or had a total value of production exceeding $10,000,000. The reported value for each commodity entry for any operation identified as an outlier then was subtracted from the "Official NASS Estimate" prior to calibration in order to create a modified NASS total. Two weights were derived using the CALJACK routine with the only difference between the two being the initial starting values for these weights. The first set of derived weights was calculated by beginning the process with all initial values set to one. The second set was calculated by beginning the calibration process using the original agricultural census weights as initial values. Next CALJACK calibrated the 1997 Census of Agriculture data to the modified NASS district totals for both acres and bushels of corn, soybeans, and cotton. Similarly modified NASS state totals were the calibration targets for acres and quantity of production for all other commodities, such as: corn silage, oats, barley, potatoes, alfalfa hay, dry beans, winter wheat; acres for snap beans, cucumbers and sugar beets; production for apples, grapes, peaches, pears, plums, sweet cherries, blueberries, and tart cherries; and, total land, all cattle, beef cows, milk cows, pigs, breeding pigs, sheep, bee colonies, hens and pullets, number of farms, number of cattle farms and number of pig farms. Following calibration, the initially removed outliers were appended to the data set in order to restore it to its complete form. This procedure preserves official

[1] An informative discussion about USDA statistics is provided in *Understanding USDA Crop Forecasts* (1999).

[2] The importance of this consideration is exemplified by inaccuracies in the 1940 census of population: 13% more African American men showed up to register for the WWII military draft than had been counted in that census.

[3] CALJACK is a modified version of a SAS macro written by Sautory of INSEE (France). It adjusts data using auxiliary information in such a way that the difference between initial and estimated weights is as small as possible. It is based on work by Deville and Särndal (1992).

state totals.

Once the agricultural survey returns have been subjected to CALJACK, the adjusted results are then post-stratified by county. This post-stratification often results in counties that do not contain sufficient data to prevent disclosure or ensure reliable production estimates. In the past such data have been suppressed in released tabulations. Of note is that while Title 13 of the United States code dictates that no data are to be published that would disclose the operations of an individual farm, the number of farms in a given county is not considered a release of confidential information. To ensure that confidentiality is maintained, historically all data have been checked with some sort of disclosure analysis, which ensures that data for an individual farm cannot be revealed directly or derived indirectly (i.e., a data user adding or subtracting a published subtotal from a published total can obtain individual farm data). The disclosure guidelines set lower limits on the number of farms that are required to be in a post-stratification group (e.g., county) before a commodity figure can be published. Publication of multiple cross-tabulations means that the actual minimum number of farms may be higher than this lower limit and may vary from county to county. Meanwhile, Perry et al. (1997) report that a threshold of 15 sample farms per county supports reliable statistical estimation. This rule is adopted here, rather than simply considering those counties actually suppressed in the census publications, because the under-coverage CALJACK-adjusted data are analyzed in this research. Meanwhile, since the total for an item for each district is published, if the value for only one county in a district is to be suppressed, then the value for at least one other county must be suppressed, too, in order to preserve confidentiality (otherwise the single suppressed value can be derived). The recommendation here is to identify the candidate counties for complementary suppression with a confidentiality algorithm, and then choose that county from this set in the district in question with the smallest item value and the largest standard error. When these two properties do not coexist for the same county, then for automating purposes that county having the smaller item value-to-standard error ratio would be appealing to choose.

Rather than remain unpublished, suppressed values could be replaced with imputed ones. Imputation also can be constrained to preserve totals, requiring that imputed values sum to the subtotal of suppressed values, by district or state, hence maintaining a "one number" census. In this context an EM (estimation-maximization)-type of algorithm can be employed. The estimation step is equivalent to a regression forecast that treats suppressed data as though they are new observations. Estimation of the regression coefficients is the maximization step. Once computed, the suppressed values' regression estimates are treated as observed values, the regression coefficients then are recomputed, and updated regression forecasts for the suppressed values are calculated. This cycle of estimation-maximization continues until parameter estimates stabilize. While the traditional EM algorithm iterates until the sum of squared terms for the imputed values goes to zero (see Navidi, 1997), constraining these imputations to sum to a given total virtually always will cause this same sum of squared terms to be non-zero. As such the convergence criterion, or any solution check, cannot be based upon these differences going to zero.

The analysis of two types of agricultural

figures can be pursued. The first is the set of totals resulting from post-stratification. The second is the set of county densities (division of a total by its corresponding county area, rendering the amount of a given agricultural commodity per 10,000 acres, say); all county areas are readily available via the NASS web page. Often ignoring county size increases nonconstant variability; hence, using densities often will yield more stable results. In contrast, using densities removes a size effect, which can result in weaker relationships being uncovered. Earlier analyses of agricultural data for Puerto Rico suggest that the use of densities is preferable. A comparison between these two types of figures is presented in the next section.

MODEL-BASED COUNTY IMPUTATIONS OF SUPPRESSED AGRICULTURAL COMMODITIES DATA

Imputation in the absence of covariates results in suppressed data being equated to the mean of the known data; this is the maximum likelihood estimate (MLE). In this context, constraining suppressed data to sum to some total results in imputations that equal the mean of the suppressed values. Of course, this mean could be a weighted average (weighted by the number of farms, say). Of note is that a geographically naive data user either will ignore counties with suppressed data (e.g., Stasny et al., 1995) or may adopt this latter solution, simply calculating the subtotal for suppressed data and dividing this quantity by the number of suppressed values. Although these estimates are useful in the absence of available redundant information, frequently they can be improved upon when useful covariates are available. Citro (1998) describes this situation as "combining data from several areas, time periods, or data sources to 'borrow strength' and improve precision." Hence, two potential sources of redundant information are the number of farms in a county, which always is disclosed, and the presence of spatial autocorrelation (i.e., counties with similar agricultural commodity levels tend to cluster together in geographic space). When production is concentrated in such a way that it is absent from a number of counties, this set of productionless counties may be viewed as coming from a different population, and hence beneficially differentiated from counties having production by using a binary indicator variable; this may be an additional useful covariate. Moderate positive spatial autocorrelation frequently is exhibited by a county map of agricultural commodities, as is attested to by the Ohio crops analyzed by Stasny et al. (1995). A hybrid imputation specification including both sources exploits all of the redundant information certain to be available to a data user. Experience shows that when reliable time series data are available, they routinely furnish a very good covariate. But because the geographic distribution of some agricultural item at an earlier point in time contains both pattern from inertia and measurement error attributable to its being for a different set of farms and/or circumstances, as well as some of its data possibly being suppressed, too, use of such temporal data is prone to error propagation that is not sufficiently well understood.

The best model-based imputations demand that the model upon which they are based is sound, and contains a minimum of specification error. If a Gaussian probability density function is to be utilized, then a power transformation may well be needed for Y. If linearity between X and Y is intrinsic (i.e, a nonlinear relationship that can be transformed to a linear one), then a power transformation

may well be needed for X. (Griffith et al., 1998) In the likely event that spatial autocorrelation is present, then an auto-regressive specification will be needed. If nonconstant variance is present, then a weighted least squares solution may be warranted. Satisfying these specific assumptions takes an analysis into the realm of nonlinear regression.

Implementation of a traditional EM-type of algorithm, itself, requires the use of nonlinear regression techniques. Consider a state that has been divided into K districts, with $n_{m.k}$ county values to be suppressed in District k, where the total number of counties in the state having suppressed (e.g., county) values is denoted by $n_m = \sum_{k=1}^{K} n_{m.k}$. Suppose the total of the suppressed values in District k is T_k; accordingly, $\sum_{k=1}^{K} T_k = T$ is the state suppressed total. Arbitrarily set one of the county weights to 1 (i.e., e^0), and let the remaining $(n_{m.k} - 1)$ county weights be denoted by $e^{u_{k.g}}$ $(g = 1, 2, ..., n_{m.k}-1; k = 1, 2, ..., K)$. Then the proportion of T_k to be allocated to county m is $w_{k,n_{m.k}} = \frac{1}{1 + \sum_{j=1}^{n_{m.k}-1} e^{u_{k.j}}}$ for the arbitrarily selected county, and $w_{k.g} = \frac{e^{u_{k.g}}}{1 + \sum_{j=1}^{n_{m.k}-1} e^{u_{k.j}}}$ for each of the remaining $g = 1, 2, ..., (n_{m.k} - 1)$ counties. The estimation problem is to calculate the $(n_{m.k} - 1)$ $u_{k.g}$ values for each of the K districts. This estimation is done iteratively: model parameters are estimated, these estimates are used to calculate imputations, the imputations are used to complete the Y data vector (i.e., suppressed values are replaced with their imputations), and then the next cycle beginning with model parameter estimation is executed. In the absence of any covariates, the MLEs of these $u_{k.g}$s equal 0. An equivalent solution involves replacing $e^{u_{k.g}}$ by $m_{k.g} \geq 0$. Nonlinear estimation utilizing this specification requires the $(n_{m.k} - 1)$ $m_{k.g}$ values to be constrained in order to insure their non-negativity, and may well encounter MLEs that become stuck at 0 (but these estimates still are legitimate).

To begin, replace the suppressed values, denoted here by a subscript m (i.e., y_m), with 0. Let X be the number of farms (X also could be defined as the quantity of variable Y in a preceding time period), denoting those for counties with observed commodity values, Y, by a subscript o (i.e., X_o), and those for counties with suppressed commodity values by X_m. Therefore, when the data are sorted so that the first n_m are the suppressed values and the next $n_o = n - n_m$ are the published values of Y,

$$\sum_{i=1}^{n} y_i =$$

$$\sum_{g=1}^{n_m} y_g + \sum_{h=n_m+1}^{n_o} y_h = \sum_{k=1}^{K} T_k + \sum_{h=n_m+1}^{n_o} y_h \ .$$

To allow for the use of either totals or densities, consider $\frac{Y_i}{D_i}$ and $\frac{X_i}{D_i}$, where the denominator term D_i is known for all i (just as X is). In this work the variable D refers either to total number of acres in a county, when densities are used, or simply 1 when totals are used. Let the indicator variable $I_{m.k.i}$ be 1 if the

commodity value is suppressed in county i of District k, and 0 otherwise. In addition, let $I_{0,i}$ be 1 if county i has 0 quantity of the agricultural item in question, and 0 otherwise. The notation employed here follows that for using indicator variables in order to implement ANOVA as a regression problem. Accordingly, in addition to the single production/nonproduction indicator variable, there are K indicator variables—one for each district—defining the districts, and the subscript i indexes the entire set of observations regardless of district membership. Since a county is contained in one and only one district, for each i only one of these indicator variables will take on a value of 1, with the (K-1) remaining values being 0. The trivariate regression equation for observation i in District g and nonlinear iteration τ in this case may be written as

$$\frac{Y_i}{D_i} = (1 - I_{m,g,i})({}_\tau\alpha + {}_\tau\beta_x \frac{X_i}{D_i} + {}_\tau\beta_0 I_{0,i}) +$$

$$\sum_{k=1}^{K} I_{m,k,i}\{({}_\tau\alpha + {}_\tau\beta_x \frac{X_i}{D_i}) -$$

$$[\frac{1 + {}_\tau w_{k,i}(T_k - n_{m,k})}{D_i}]\} + \epsilon_i , \qquad (1)$$

where the nonlinear regression parameter estimates ${}_0\alpha$ and ${}_0\beta_j$ (j = x, 0) can be initialized with their counterpart estimates obtained from regressing $\frac{Y_o}{D_o}$ on $\frac{X_o}{D_o}$ and $I_{0,o}$ (recall the subscript "o" denotes unsuppressed or observed), $w_{k,i} = 0$ for all unsuppressed data, and ϵ is the error term to which most of the modeling assumptions pertain. A convenient initial value for ${}_0w_{k,i}$ is furnished by $m_{k,g} = 1$ (i.e., the uniform distribution); as such, the ${}_0w_{k,i}$ values initialize imputation with the means of the suppressed values. Writing the suppressed data term in equation (1) as $\frac{1 + {}_\tau w_{k,i}(T_k - n_{m,k})}{D_i}$ ensures that a minimum value of 1 is estimated (the 1 appearing in the left-hand term of the numerator) for each of the $n_{m,k}$ counties (hence, $n_{m,k}$ must be subtracted from T_k), uses the ${}_\tau w_{k,i}$ weights (which all are non-negative and sum to 1) to allocate a percentage of the remaining total of the suppressed values (i.e., $T_k - n_{m,k}$) to each candidate county at iteration τ, expresses the estimation in terms of densities (the D_i term in the denominator), and avoids the need to compute a back-transformation since $w_{k,i}$ is being directly estimated[4]. The iterative estimation occurs as described in Navidi (1997), employs a standard nonlinear regression algorithm (e.g., steepest descent, Newton-Raphson, Marquardt), is based upon the partial derivatives of the regression function with respect to each of the parameters to be estimated, and at each iteration τ involves simultaneous estimation of $(n_m - k + 3)$ parameters, namely α, β_x, β_0, and $u_{k,g}$ or $m_{k,g}$ (g = 1, 2, ...,$n_{m,k}$-1; k = 1, 2, ..., K); k is subtracted because one parameter in each district, or a single one for the state when only a statewide constraint is imposed, is arbitrarily set to unity. Applying a variance stabilizing and/or normality aspiring power transformation to the data converts equation (1) to

$$(\frac{Y_i}{D_i} + \delta)^\gamma = (1 - I_{m,g,i})[{}_\tau\alpha + {}_\tau\beta_x(\frac{X_i}{D_i} + \theta)^\eta +$$

---

[4] In general the term is $\frac{a + {}_\tau w_{k,i}(T_k - a \times K)}{D_i}$, $0 \le a \le \frac{T_k}{K}$. Because absence of production is reported in census publications, setting a = 1 as a minimum seems reasonable.

$$_{\tau}\beta_0 I_{0,i}] + \sum_{k=1}^{K} I_{m,k,i}\{[_{\tau}\alpha + {}_{\tau}\beta_x(\frac{X_i}{D_i}+\theta)^{\eta}] -$$

$$[\frac{1 + {}_{\tau}w_{k,i}(T_k - n_{m,k})}{D_i} + \delta]^{\gamma}\} + \epsilon_i \ , \quad (2)$$

where δ and γ are selected to optimize $\frac{Y}{D}$'s conforming to a normal frequency distribution, and θ and η are selected to optimize the linear relationship displayed between the power transformed versions of $\frac{X}{D}$ and $\frac{Y}{D}$. Meanwhile, θ and η respectively often will take on the same values as δ and γ, although they can differ since their purpose is to maximize the linearity of the relationship displayed between the power-transformed versions of $\frac{Y}{D}$ and $\frac{X}{D}$. If I_0 is not a useful covariate then β_0 should be set to 0, reducing equations (1) and (2) to bivariate forms. In part the usefulness of I_0 will depend upon the relative number of counties in a state in which a given agricultural item is absent. This type of indicator variable also could be used to differentiate outliers.

Rather than a trivariate regression specification—if no highly correlated covariates are available (e.g., the weighted number of farms is a poor predictor of $\frac{Y}{D}$)—the modeling can employ a pure spatial auto-Gaussian[5] specification of the form

$$\frac{Y_i}{D_i} = {}_{\tau}\rho\{\sum_{j=1}^{n_o} c_{ij}\frac{Y_j}{D_j} + \sum_{k=1}^{K}\sum_{h=1}^{n_{m,k}} c_{ih,k} \times$$

$$[\frac{1 + {}_{\tau}w_{k,h}(T_k - n_{m,k})}{D_h}]\}/\sum_{j=1}^{n} c_{ij} +$$

$$(1 - {}_{\tau}\rho)_{\tau}\alpha - \sum_{k=1}^{K} I_{m,k,i}\times$$

$$[\frac{1 + {}_{\tau}w_{k,i}(T_k - n_{m,k})}{D_i}] + \epsilon_i \ , \quad (3)$$

where ρ is a spatial autocorrelation parameter, and c_{ij} is the binary (0 or 1) entry of the geographic connectivity or weights matrix[6]. Applying a variance stabilizing and/or normality aspiring power transformation to the data converts equation (3) to

$$(\frac{Y_i}{D_i} + \delta)^{\gamma} =$$

$${}_{\tau}\rho\{\sum_{j=1}^{n_o} c_{ij}(\frac{Y_j}{D_j} + \delta)^{\gamma} + \sum_{k=1}^{K}\sum_{h=1}^{n_{m,k}} c_{ih,k} \times$$

$$[\frac{1 + {}_{\tau}w_{k,h}(T_k - n_{m,k})}{D_h} + \delta]^{\gamma}\}/\sum_{j=1}^{n} c_{ij} + (1 - {}_{\tau}\rho)_{\tau}\alpha$$

$$- \sum_{k=1}^{K} I_{m,k,i}[\frac{1 + {}_{\tau}w_{k,i}(T_k - n_{m,k})}{D_i} + \delta]^{\gamma} + \epsilon_i \ . \quad (4)$$

Because $\hat{\rho}$ very often falls between 0.4 and

[5] The first derivative with respect to ρ is nonstandard, and must be determined externally in order for SAS to have it in its correct analytical form.

[6] A geographic connectivity or weights matrix is an n-by-n binary, 0-1 matrix C whose row and column labels are the same sequence of the counties that is defined by the i subscript attached to Y. Cell entry c_{ij} is 1 if two counties share a common boundary, and 0 otherwise, with c_{ii} being set to 0. This matrix defines the pairs of counties whose values are directly spatially correlated. Of note is that more sophisticated measure of c_{ij} can be used (e.g., interpoint distances, percent of common boundary).

0.6, $_{\tau=0}\rho = 0.5$ is a good starting value for initiating the nonlinear estimation of either equation (3) or (4).

The hybrid model that integrates equations (1) and (3) is

$$\frac{Y_i}{D_i} = {}_{\tau}\rho\{\sum_{j=1}^{n_o} c_{ij}\frac{Y_j}{D_j} + \sum_{k=1}^{K}\sum_{h=1}^{n_{m,k}} c_{ih,k} \times$$

$$[\frac{1 + {}_{\tau}w_{k,h}(T_k - n_{m,k})}{D_h}]\}/\sum_{j=1}^{n} c_{ij} +$$

$$(1 - {}_{\tau}\rho)_{\tau}\alpha +$$

$${}_{\tau}\beta_x\{\frac{X_i}{D_i} - {}_{\tau}\rho\sum_{j=1}^{n} c_{ij}\frac{X_j}{D_j}/\sum_{j=1}^{n} c_{ij}\} +$$

$${}_{\tau}\beta_0\{I_{0,i} - {}_{\tau}\rho\sum_{j=1}^{n} c_{ij}\, I_{0,j}/\sum_{j=1}^{n} c_{ij}\} -$$

$$\sum_{k=1}^{K} I_{m,k,i}[\frac{1 + {}_{\tau}w_{k,i}(T_k - n_{m,k})}{D_i}] + \epsilon_i \quad ,(5)$$

whereas the hybrid model that integrates equations (2) and (4) is

$$(\frac{Y_i}{D_i} + \delta)^{\gamma} =$$

$${}_{\tau}\rho\{\sum_{j=1}^{n_o} c_{ij}(\frac{Y_j}{D_j} + \delta)^{\gamma} + \sum_{k=1}^{K}\sum_{h=1}^{n_{m,k}} c_{ih,k} \times$$

$$[\frac{1 + {}_{\tau}w_{k,h}(T_k - n_{m,k})}{D_h} + \delta]^{\gamma}\}/\sum_{j=1}^{n} c_{ij} +$$

$$(1 - {}_{\tau}\rho)_{\tau}\alpha + {}_{\tau}\beta_x\{(\frac{X_i}{D_i} + \theta)^{\eta} -$$

$${}_{\tau}\rho\sum_{j=1}^{n} c_{ij}(\frac{X_j}{D_j} + \theta)^{\eta}/\sum_{j=1}^{n} c_{ij}\} +$$

$${}_{\tau}\beta_0\{I_{0,i} - {}_{\tau}\rho\sum_{j=1}^{n} c_{ij}\, I_{0,j}/\sum_{j=1}^{n} c_{ij}\} -$$

$$\sum_{k=1}^{K} I_{m,k,i}[\frac{1 + {}_{\tau}w_{k,i}(T_k - n_{m,k})}{D_i} + \delta]^{\gamma} + \epsilon_i \quad .(6)$$

Of note is that if $_{\tau}\rho = 0$ then equations (5) and (6) respectively reduce to equations (1) and (2), and if $\delta = \theta = 0$ and $\gamma = \eta = 1$ then equations (5) and (6) respectively reduce to equations (3) and (4).

Experience suggests that when $n_{m,k}$ is more than 2 or 3, nonlinear regression convergence is easier to achieve for equations (3)-(6) using the $w_{k,g}$ specification based upon $m_{k,g}$ rather than $e^{u_{k,g}}$. Furthermore, experience confirms that an optimal pair of values for δ and γ, as well as θ and η, is not guaranteed to exist, and that frequently $\delta > 0$ (and often approximately 0) when densities are used .

THE 1997 MICHIGAN DATA ANALYSIS

The major agricultural commodities in Michigan are corn and soybeans, both measured in acres and in bushels, beef cattle, and dairy cattle. The state of Michigan is partitioned into n=83 counties, which are regionalized into K=9 districts. The data were collected for 42,084 individual farm operations (which translate to 51,044 CALJACK-equivalent

Table 1: Post-stratification of Michigan Farm Types by County, and Affiliated County Suppression Statistics

Agricultural commodity	Corn	Soybeans	Beef cattle	Milk cows
number of farms	18,286	13,970	8,378	4,119
number of counties with 0 farms	6	19	0	5
% of commodity suppressed	acres: 0.15 bushels: 0.11	acres: 0.23 bushels: 0.14	0.86	4.02
number of counties with 1-14 farms	District 10: 8 District 30: 4 District 40: 1	District 10: 6 District 20: 4 District 30: 7 District 50: 3	5	24

Table 2: Districts, Combined Districts, or State Totals to Which Suppressed Values must Sum, for Michigan

Geographic aggregate	Corn	Soybeans	Beef cattle	Milk cows
District 10	acres: 795 bushels: 69,078	acres: 3,631 bushels: 81,226		
District 20				
District 30	acres: 2,232 bushels: 202,096			
District 40	acres: 9,243‡ bushels: 888,816‡			
District 50		acres: 670 bushels: 20,156		
state			1,012	12,059

‡Includes the figure for arbitrarily selected Mason County, which serves as a complementary county here.
Note: Data for shaded cells are not relevant

farms[7]) and recorded for the county of each farm's headquarters. Only NASS district totals for corn and soybeans were considered reliable enough for calibration purposes—these are the figures to which CALJACK calibrated census results. The NASS state totals were considered better calibration benchmarks for all other commodities. Post-stratification by county reveals the tabulations reported in Table 1. Totals for the suppressed values (i.e., figures for counties having $0 < n_{f,i} \leq 14$, where $n_{f,i}$ denotes the number of farms for county i) appear in Table 2. These are the totals, T_k, to which suppressed county commodity estimates must sum by district, or the totals, T, to which suppressed county commodity estimates must sum for the state.

Of note is that the spatial autocorrelation displayed by the geographic distribution of corn acres and production (in bushels) is almost completely accounted for by spatial autocorrelation latent in the number of farms, once an appropriate power transformation has been applied to this predictor variable. Accordingly, the hybrid model exhibits only a slight improvement over the bivariate regression model. Modest spatial autocorrelation remains in soybean acres and bushels after these variables have been regressed on the number of farms. But a trivariate regression model is needed here for county totals (but not for densities), as the indicator variable for 0-production counties is significant. The geographic distributions of both beef cattle and milk cows continue to display weak-to-moderate spatial autocorrelation after these variables have been regressed on the number of farms. And, the milk cows variable contains a spatial-and-aspatial outlier.

[7]Virtually all CALJACK equivalent-farm counts will be inflated versions of the input numbers of farms, since mostly CALJACK is making weighting adjustments for failure to survey not-on-the-list frame farms.

Field Corn for Grain Harvested

Compared with its untransformed version, application of a Box-Cox power transformation to the corn data reveals that both acre and bushel totals and densities of corn by county better conform to a normal frequency distribution when $\delta = 0$ and $\gamma = ¼$: the Shapiro-Wilk[8] (S-W) statistic improves from $0.78154^{\dagger\dagger\dagger}$ to $0.92850^{\dagger\dagger\dagger 9}$ for total acres, from $0.79718^{\dagger\dagger\dagger}$ to $0.92880^{\dagger\dagger\dagger}$ for density of acres, from $0.75760^{\dagger\dagger\dagger}$ to $0.93294^{\dagger\dagger\dagger}$ for total bushels, and from $0.77180^{\dagger\dagger\dagger}$ to $0.93409^{\dagger\dagger\dagger}$ for density of bushels. For acres of corn the Moran Coefficients (MCs) measuring spatial autocorrelation are: $0.64387^{\dagger\dagger\dagger}$ and $0.65392^{\dagger\dagger\dagger}$, respectively, for the raw totals and densities data, and $0.77610^{\dagger\dagger\dagger}$ and $0.77565^{\dagger\dagger\dagger}$, respectively, for the power-transformed totals and densities data. For bushels of corn the MCs are: $0.58540^{\dagger\dagger\dagger}$ and $0.59220^{\dagger\dagger\dagger}$, respectively, for the raw totals and densities data, and $0.77530^{\dagger\dagger\dagger}$ and $0.77391^{\dagger\dagger\dagger}$, respectively, for the power-transformed totals and densities data. In other words, moderate levels of positive spatial autocorrelation are present in these data. With only six counties having no corn production, b_0 is found not to be significant, indicating the adoption of a bivariate regression equation specification.

Restricting attention to the power-transformed data, and setting $\theta = 0$ and $\eta = ⅓$ for the number of farms[10], estimation of the bivariate regression equation corresponding to equation (2) but for the complete data set yields

totals: $a = 0.13169$, $b_x = 1.99150$,
$R^2 = 0.982$, S-W = 0.97514,
MC = 0.06414,

and

densities: $a = 0.00379$, $b_x = 5.84552$,
$R^2 = 0.982$, S-W = 0.98202,
MC = 0.05911.

In other words, the geographic distribution of number of corn farms accounts for most of the variance in the geographic distribution of acres of corn, as well as a large portion of the spatial autocorrelation displayed by the geographic distribution of acres of corn. Considerable strength is available to borrow from if model-based estimation for small geographic areas is to be undertaken. Of note is that the residuals contain only a trace amount of spatial autocorrelation. For initial quick model identification purposes, a user faced with suppressed data can get an estimate of the MC in this case by regressing Y_o on X_o and then assigning the suppressed value residual (Y_m - a - b X_m) a value of 0—recall that this term iterates to 0 with the traditional EM algorithm. Accordingly, for these corn data a user would obtain

totals: $a = 0.16346$, $b_x = 1.98669$,
$R^2 = 0.979$, S-W = 0.97261,
MC = 0.06902,

and

densities: $a = 0.00223$, $b_x = 5.86183$,
$R^2 = 0.977$, S-W = 0.97376,

[8]The S-W statistic is based upon a generalized least squares regression of ordered sample values on normal scores, and hence may be thought of as the squared correlation coefficient between these ordered sample values and a set of values that is approximately proportional to the normal scores. As such it is a measure of the straightness of the corresponding normal probability plot. Accordingly, small values indicate departure from normality.

[9]††† denotes a significant difference at the 1% level; †† denotes a significant difference at the 5% level; † denotes a significant difference at the 10% level.

[10]County figures from either the 1992 Census of Agriculture, or the 1997-98 NASS report contain suppressed values. The relevant census table includes values for 70 of the 83 counties, whereas the relevant NASS table includes values for 52 of the 83 counties.

MC = 0.04766,

which do not differ greatly from the corresponding complete data results, allowing a user to make sound decisions about which suppressed data equation to implement. When compared with their univariate counterparts, these two MC values have dramatically shrunk toward zero, which is indicative of much of the spatial autocorrelation contained in the original agricultural commodity variables being accounted for by the spatial autocorrelation contained in the farm counts variable.

Meanwhile, estimating a pure spatial autoregressive model, corresponding to equation (4) but for the complete data set, yields

totals: $a = 9.67347$, $\hat{\rho} = 0.89685$,
pseudo-$R^2 = 0.855$, S-W = 0.98496,
MC = -0.04655,

and

densities: $a = 0.40177$, $\hat{\rho} = 0.90358$,
pseudo-$R^2 = 0.865$, S-W = 0.97551,
MC = -0.04260 .

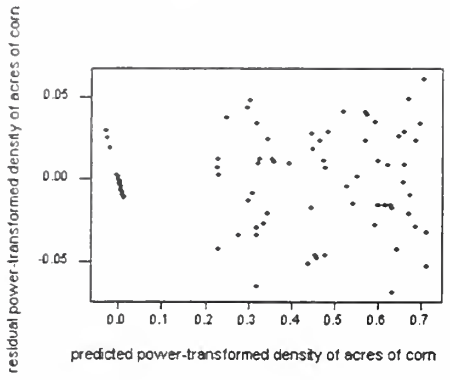
These results reveal that strong positive spatial autocorrelation is present in these corn data—correlation that is essentially completely accounted for by a SAR model specification (the MC value is approximately equal to its OLS expected value counterpart), that nearby values account for roughly 86% of the variation in Y, and that the SAR residuals conform to a normal frequency distribution. In other words, again considerable strength is available to borrow from if model-based estimation for small geographic areas is to be undertaken.

Estimating equations (2) and (4) as well as the hybrid specification, corresponding to equation (6), in the presence of 14 suppressed data values, which account for only 0.41% of the total corn acreage, yields suppressed data estimates with the properties summarized in Table 3. The mean squared prediction error (MSPE) is given by $\frac{\sum_{i=1}^{n_m} (y_m - \hat{y}_m)^2}{n_m}$, with $\{\hat{y}_m\}$ being the set of predicted values rendered by equations (2), (4), (6), or, for the weighted mean, $\frac{n_{i,k} \times T_k}{\sum_{i=1}^{k} n_{i,k}}$. These results show that for this case borrowing strength from

Table 3: Diagnostic Statistics for Suppressed Data Estimates, Acres of Corn

Estimator	Mean squared prediction error (MSPE)	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
totals			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	90,262	0.93093†††	*
equation (2)	61,933	0.98004	0.01247
equation (4)	331,9961	0.97042	0.10066
equation (6) [$\hat{\rho} = 0.27231$]	41,030	0.98002	0.01176
densities			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	90,262	0.93222†††	*
equation (2)	63,757	0.98007	0.01298
equation (4)	301,1976	0.96074††	0.09647
equation (6) [$\hat{\rho} = 0.21562$]	38,155	0.97751	0.01254
Note: * denotes RSSE cannot be calculated.			

Figure 1: Standard homogeneity of variance ($y - \hat{y}$) versus $\hat{y}$ regression plot. Of note is that the SAR-imputed values cluster together, forming a straight line in the left-hand portion of the graph. The right-hand portion of the graph essentially suggests constant variance for the SAR model.



nearby county values coupled with the number of farms as a covariate dramatically improves upon the weighted mean of the suppressed values as the estimator. The hybrid model produces the best estimates, the bivariate regression model the second-best estimates, and the weighted district means produce the third-best estimates. There is little difference between estimates based upon totals and those based upon densities. Of note is that two markedly wild imputations caused the pure SAR model to have poor performance here. Model diagnostics for the hybrid model are very good (see Figure 1, which displays the traditional homogeneity of variance residual plot). The imputed values generated here, after rounding to integers, are reported in Table 4. Of note is that eight of these counties are located in the upper peninsula, which constitutes District 10. Unconstrained estimates give residual totals for Districts 10, 30, and 40 of, respectively, 814, 1,285, and 9,864 for the totals data, and 665, 1,203, and 10,432 for the densities data—rather than the corresponding values, reported previously, of 795, 2,232, and 9,243 needed for a one-number census. In other words, employing auxiliary information supplied by totals dampens the over- and under-estimations of suppressed values generated by a model like equation (6). These discrepancies emphasize the need to employ constrained EM-type estimation if the notion of a one-number census is to be preserved. In terms of percentage deviation from the post-stratified county values, the most problematic case is the one with the smallest NASS value, whereas the best case is the one with the largest value (which incidentally is the case that was arbitrarily selected to ensure confidentiality in District 40).

Table 4: Suppressed Data Estimates for Michigan Counties, Acres of Corn

County	Estimate based on totals	Estimate based on densities
Cheboygan	867	858
Chippewa	24	16
Crawford	61	59
Dickinson	435	475
Houghton	70	63
Iron	19	17
Lake	418	404
Luce	82	78
Mackinac	23	21
Marquette	25	21
Mason	8,825	8,839
Montmorency	849	874
Oscoda	455	441
Schoolcraft	117	105

Table 5: Diagnostic Statistics for Suppressed Data Estimates, Bushels of Corn

Estimator	Mean squared prediction error (MSPE)‡	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
totals			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	10.1	0.93482†††	*
equation (2)	7.3	0.98263	0.01642
equation (4)	337.6	0.97142	0.10416
equation (6) [$\hat{\rho} = 0.15097$]	5.4	0.98106	0.01567
densities			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	10.1	0.93711†††	*
equation (2)	7.5	0.97187	0.01721
equation (4)	305.7	0.96749	0.10127
equation (6) [$\hat{\rho} = 0.18591$]	4.1	0.96952	0.00548

‡These figures need to be multiplied by $\times 10^8$.
Note: * denotes RSSE cannot be calculated.

Repeating this exercise for bushels of corn, whose 14 suppressed values account for only 0.44% of the total number of bushels, produces the results tabulated in Table 5. These results also show that borrowing strength from the number of corn farms as a covariate dramatically improves upon the mean of the suppressed values as the estimator. Again in this case latent spatial autocorrelation alone does not produce improved estimates.[11] This finding more than likely is attributable to the presence of a nonconstant mean beyond the one captured by a spatial autoregressive term. The hybrid model achieves a lower RSSE and an acceptable S-W statistic, especially when estimation is based upon densities, implying a sounder model. As in the case of acres of corn harvested, on some, if not most, indices all three model-based estimates are better than the weighted mean of the suppressed values. The imputed values generated here, after rounding to integers, are reported in Table 6. In terms of percentage deviation from the post-stratified county values, the worst case by far again is for Marquette (the smallest value), whereas the best case is for Mason (with the largest value, and being the arbitrarily suppressed, complementary county value).

Soybeans Harvested

Analysis of soybean data reveals that both acres and bushels of soybeans harvested, by county, better conform to a normal frequency distribution when $\delta = 0$ and $\gamma = \frac{1}{5}$: the Shapiro-Wilk (S-W) statistic improves from 0.71990††† to 0.87425††† for total acres and from 0.73194††† to 0.86775††† for density of acres, and from 0.70721††† to 0.87147††† for total bushels and from 0.72210††† to 0.86585††† for density of bushels. Neither of these power-transformed values suggests adequate conformity with a normal distribution. For acres of soybeans the MCs measuring spatial autocorrelation are: 0.58900††† and 0.63301†††, respectively, for the raw totals and densities data, and 0.82664††† and 0.82346†††, respectively, for the power-transformed totals and densities data. For bushels of soybeans

[11]These results still are better than their unconstrained traditional EM counterparts, whose statewide total is 1% too little, and whose district totals differ as follows: the District 10 total is underpredicted by 20.3%, the District 30 total is underpredicted by 51.2%, and District 40 total is overpredicted by 14.6%.

Table 6: Suppressed Data Estimates for Michigan Counties, Bushels of Corn

County	Estimate based on totals	Estimate based on densities
Cheboygan	88,504	78,335
Chippewa	1,205	565
Crawford	3,794	3,704
Dickinson	41,262	46,749
Houghton	6,983	5,965
Iron	1,389	1,056
Lake	33,111	31,190
Luce	6,333	5,518
Mackinac	1,207	877
Marquette	1,291	850
Mason	855,705	857,626
Montmorency	76,229	83,930
Oscoda	33,508	36,127
Schoolcraft	9,408	7,495

the MCs are: $0.55506^{\dagger\dagger\dagger}$ and $0.60013^{\dagger\dagger\dagger}$, respectively, for the raw totals and densities data, and $0.83641^{\dagger\dagger\dagger}$ and $0.83251^{\dagger\dagger\dagger}$, respectively, for the power-transformed totals and densities data. In other words, moderate levels of positive spatial autocorrelation are present in these data. With 19 counties having no soybean production, b_0 is found to be significant for the totals data, indicating the adoption of a trivariate regression equation specification.

Restricting attention to the power-transformed densities data, and employing the Box-Cox analysis results of $\theta = 0$ and $\eta = \frac{1}{4}$ (note, $\eta = \frac{1}{5}$ for the totals data), estimation of equation (2) yields

$$a = 0.00167,\ b_x = 3.74283,\ b_0 = -0.00167\ (s_{b_0} = 0.01023),\ R^2 = 0.990,$$
$$\text{S-W} = 0.93783^{\dagger\dagger\dagger},\ \text{MC} = 0.11491.$$

In other words, the geographic distribution of density of soybean farms accounts for virtually all of the variance in the geographic distribution of the density of acres of soybeans, as well as much of the spatial autocorrelation displayed by the geographic distribution of acres of soybeans. Considerable strength is available to borrow from if model-based estimation for small geographic areas is to be undertaken. In addition, whereas 0 totals appear to be from a different population, 0 densities do not. Approximation of the MC provided by setting $(Y_m - a - b_x X_m)$ to 0 yields

$$a = 0.00289,\ b_x = 3.74050,\ R^2 = 0.993,$$
$$\text{S-W} = 0.86508^{\dagger\dagger\dagger},\ \text{MC} = 0.05094,$$

which does not differ greatly from the complete data results, again allowing a user to make reasonable decisions about which suppressed data equation to implement. Of note is that the spatial autocorrelation index differs noticeably when totals are used because the zero residuals for suppressed value estimates are confused with the zero residuals for the counties in which soybean production is absent.

Meanwhile, estimating a pure spatial autoregressive model, corresponding to equation (4) but for the complete data set, yields

$$\text{totals: } a = 4.49086,\ \hat{\rho} = 0.92494,$$
$$\text{pseudo-}R^2 = 0.896,\ \text{S-W} = 0.98148,$$

MC = -0.05758 ,

and

densities: a = 0.35431, $\hat{\rho}$ = 0.92612,
pseudo-R^2 = 0.898, S-W = 0.98325,
MC = -0.04893.

These results reveal that strong positive spatial autocorrelation is present in these soybean data—correlation that is essentially completely accounted for by an SAR model specification (the MC value is approximately equal to its OLS expected value counterpart of −0.02228). In this case nearby values account for roughly 90% of the variation in Y, and the SAR residuals conform to a normal frequency distribution. In other words, as with corn, considerable strength is available to borrow from if model-based estimation for small geographic areas is to be undertaken.

Although reliable district totals are available for soybean acreage, a paucity of farms in Districts 10, 20 and 30 prevented CALJACK from converging. These three coterminous districts were pooled in order to successfully attain convergence with CALJACK. Hence the sum of these three district totals is used as a constraining total in the estimation process. This pooling of district figures demonstrates the flexibility of the small area estimation procedure being described here. The pooling of districts reduces the amount of constraint information—in this case from 4 totals to 2 totals—and increases the variability of the estimates. Estimating equations (2) and (4) as well as the hybrid specification, corresponding to equation (6), in the presence of 20 suppressed data values, which account for only 0.23% of the total soybean acreage, yields suppressed data estimates with the properties reported in Table 7. These results show that for this case borrowing strength from the number of soybean farms as a covariate improves upon the weighted mean of the suppressed values as the estimator. Considering the totals data, the hybrid model produces the best estimates, but only marginally better ones than the trivariate regression model, which produces the second-best estimates. Unfortunately, on average the densities data models do not clearly outperform the weighted means solution. Perhaps forcing predicted values to 0 for nonproducing counties is preferable even if b_0 is nonsignificant. Model diagnostics for the hybrid model are somewhat disappointing, though.

Table 7: Diagnostic Statistics for Suppressed Data Estimates, Acres of Soybeans

Estimator	Mean squared prediction error (MSPE)	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
totals			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	20,988.4	0.92156†††	*
equation (2)	20,797.7	0.86250†††	0.00422
equation (4)	85,857.3	0.94040†††	0.03941
equation (6) [$\hat{\rho}$ = 0.04672]	20,596.5	0.86353†††	0.00422
densities			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	20,988.4	0.92321†††	*
equation (2)	21,792.7	0.87030†††	0.00464
equation (4)	96,986.4	0.93937†††	0.03898
equation (6) [$\hat{\rho}$ = 0.15019]	21,715.8	0.87811†††	0.00457
Note: * denotes RSSE cannot be calculated.			

Table 8: Suppressed Data Estimates for Michigan Counties, Acres of Soybeans

County	Estimate based on totals	Estimate based on densities
Alcona	467	412
Alger	21	44
Alpena	400	395
Cheboygan	21	42
Clare	443	394
Delta	116	137
Grand Traverse	657	621
Mackinac	21	42
Manistee	20	41
Marquette	21	37
Mecosta	166	187
Menominee	116	143
Missaukee	21	44
Montmorency	178	213
Ogemaw	306	299
Osceola	62	90
Otsego	177	217
Presque Isle	953	764
Schoolcraft	116	136
Wexford	16	44

The imputed values generated here, after rounding to integers, are reported in Table 8. Unconstrained estimates give residual totals for the aggregation of Districts 10, 20 and 30, and for District 50 of, respectively, 2947 and 584 based upon totals, and 4554 and 935 based upon densities—rather than the corresponding values, reported previously, of 3631 and 670 needed for a one-number census. Again, employing auxiliary information supplied by totals dampens the over- and underestimations of suppressed values generated by a model like equation (6). In terms of percentage deviation from the post-stratified county values, the most problematic case is the one with the smallest NASS value, whereas the best is for the median of the smaller set of counties. Of note is that latent spatial autocorrelation is almost completely accounted for by the geographic distribution of soybean farms when the 20 suppressed values are estimated. The impact of so many estimated values seems to have little detectable affect upon the pure SAR results.

Repeating this exercise for bushels of soybeans, whose 20 suppressed values account for only 0.14% of the total number of bushels of soybeans, produces the results tabulated in Table 9. These results also show that borrowing strength from the number of farms as a covariate improves upon the weighted mean of the suppressed values as the estimator for the totals data, but not for the densities data, and that spatial autocorrelation alone does not furnish as much redundant information as does a good covariate. In this case the hybrid model is not clearly superior to its EM counterpart, primarily because the spatial autocorrelation latent in the geographic distribution of bushels of soybeans is completely accounted for by the spatial autocorrelation contained in the geographic distribution of farms producing soybeans. The imputed values generated here, after rounding to integers, are reported in Table 10. Unconstrained estimates give residual totals for the aggregation of Districts 10, 20 and 30, and for District 50 of, respectively, 73,356 and 14,054 based

Table 9: Diagnostic Statistics for Suppressed Data Estimates, Bushels of Soybeans

Estimator	Mean squared prediction error (MSPE)	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
totals			
suppressed values weighted means (i.e., $n_{m_i}T_k/n_{m_k}$)	16,152,637	0.92175†††	*
equation (2)	15,776,927	0.89662†††	0.00464
equation (4)	66,838,055	0.94235†††	0.04350
equation (6) [$\hat{\rho}$ = -0.03676]	15,876,480	0.89733†††	0.00464
densities			
suppressed values weighted means (i.e., $n_{m_i}T_k/n_{m_k}$)	16,153,679	0.92410†††	*
equation (2)	17,787,034	0.90340†††	0.00508
equation (4)	73,077,741	0.93926†††	0.04309
equation (6) [$\hat{\rho}$ = 0.05515]	17,706,362	0.90437†††	0.00507

Note: * denotes RSSE cannot be calculated.

upon totals, and 112,355 and 23,254 based upon densities—rather than the corresponding values, reported previously, of 81,225 and 20,156. This would be roughly a 14% undercount. As before, employing auxiliary information supplied by totals dampens the over- and under-estimations of suppressed values generated by a model like equation (6).

In terms of percentage deviation from the post-stratified county values, the worst case by far again is for the county having the smallest number of bushels of soybean production, whereas the best case is for the county having

Table 10: Suppressed Data Estimates for Michigan Counties, Bushels of Soybeans

County	Estimate based on totals	Estimate based on densities
Alcona	11,222	9,941
Alger	251	588
Alpena	8,901	9,147
Cheboygan	251	627
Clare	14,900	13,149
Delta	2,157	2,353
Grand Traverse	15,523	15,486
Mackinac	251	555
Manistee	260	695
Marquette	251	415
Mecosta	4,179	5,047
Menominee	2,157	2,488
Missaukee	249	713
Montmorency	3,554	4,574
Ogemaw	7,171	7,339
Osceola	1,077	1,959
Otsego	3,555	4,724
Presque Isle	23,063	18,534
Schoolcraft	2,158	2,342
Wexford	251	704

the largest amount.

Head of Beef Cattle

Analysis of beef cattle data reveals that head counts by county better conform to a normal frequency distribution when $\delta = -7$ and $\gamma = ½$

for totals, and $\delta = 0$ and $\gamma = ½$ for densities. Comparing the raw data with the Box-Cox transformed data, the Shapiro-Wilk (S-W) statistic improves from $0.91943^{\dagger\dagger\dagger}$ to 0.97371 for total head, and from $0.91435^{\dagger\dagger\dagger}$ to 0.97984 for density of head, suggesting adequate conformity with a normal distribution. The accompanying MCs measuring spatial autocorrelation are: $0.59397^{\dagger\dagger\dagger}$ and $0.59901^{\dagger\dagger\dagger}$, respectively, for the raw totals and densities data, and $0.53648^{\dagger\dagger\dagger}$ and $0.59098^{\dagger\dagger\dagger}$, respectively, for the power-transformed totals and densities data. In other words, the commonly found moderate level of positive spatial autocorrelation is present in these data. All counties have at least some beef cattle, indicating the adoption of a bivariate regression equation specification.

Restricting attention to the power-transformed density data, and setting $\theta = 0$ (rather than -4, the value needed if totals data are employed) and $\eta = ½$, estimation of equation (2) yields

$$a = 0.00499,\ b_x = 3.41277,\ R^2 = 0.883,$$
$$\text{S-W} = 0.95742^{\dagger\dagger},\ \text{MC} = 0.24540^{\dagger\dagger\dagger}.$$

In other words, the geographic distribution of density of beef cattle farms accounts for nearly 88% of the variance in the geographic distribution of head of beef cattle, as well as about 50% of the spatial autocorrelation displayed by this geographic distribution. Considerable strength is available to borrow from if model-based estimation for small geographic areas is to be undertaken. Approximation of the MC provided by setting $(Y_m - a - b_x X_m)$ to 0 yields

$$a = 0.00575,\ b_x = 3.36978,\ R^2 = 0.864,$$
$$\text{S-W} = 0.95853^{\dagger\dagger},\ \text{MC} = 0.24463^{\dagger\dagger\dagger},$$

which does not differ very much from the complete data results, once more allowing a user to make reasonable decisions about which suppressed data equation to implement.

Meanwhile, estimating a pure spatial autoregressive model, corresponding to equation (4) but for the complete data set, yields

totals: $a = 6.55044,\ \hat{\rho} = 0.69839,$
pseudo-$R^2 = 0.517$, S-W = 0.98338,
MC = -0.01352 ,

and

densities: $a = 0.05397,\ \hat{\rho} = 0.75330,$
pseudo-$R^2 = 0.606$, S-W = 0.97635,
MC = 0.00920.

These results reveal that moderate positive spatial autocorrelation is present in these beef cattle data—correlation that is essentially completely accounted for by an SAR model specification (the MC value is approximately equal to its OLS expected value counterpart). In this case nearby values account for roughly half of the variation in Y, and the SAR residuals conform to a normal frequency distribution. In other words, as with corn and soybeans, considerable redundant information is available to borrow from if model-based estimation for small geographic areas is to be undertaken.

Because district totals were not considered to be sufficiently reliable to calibrate beef cattle, CALJACK calibration was to the state total. Accordingly, the suppressed county values estimated here (coming from nearly all of the districts) are constrained to this single state total. One advantage of this weaker constraint is that for each of the two districts having only a single county value suppressed, additional complementary counties need not have their values suppressed. Estimating equations (2) and (4) as well as the hybrid specification, corresponding to equation (6), in the presence

Table 11: Diagnostic Statistics for Suppressed Data Estimates, Head of Beef Cattle

Estimator	Mean squared prediction error (MSPE)	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
totals			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	6,745.8	0.96342†	*
equation (2)	5,594.2	0.96486†	0.09763
equation (4)	11,363.6	0.97023	0.42589
equation (6) [$\hat{\rho} = 0.38288$]	3,133.8	0.96798	0.08600
densities			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	6,745.8	0.97101	*
equation (2)	8,155.8	0.95840††	0.09213
equation (4)	29,048.0	0.97285	0.34770
equation (6) [$\hat{\rho} = 0.34979$]	6,329.2	0.96161†	0.08271
Note: * denotes RSSE cannot be calculated.			

of 5 suppressed data values, which account for merely 0.14% of the total number of beef cattle, yields suppressed data estimates with the properties tabulated in Table 11. These results show that for this case borrowing strength from the number of beef cattle farms as a covariate can improve upon the weighted mean of the suppressed values as the estimator. The hybrid model produces the best estimates, ones that are dramatically better than those rendered by the bivariate regression model, which produces the second-best estimates for the totals data. The failure for number of farms to be as strong a covariate

Table 12: Suppressed Data Estimates for Michigan Counties, Head of Beef Cattle

County	Estimate based on totals	Estimate based on densities
Alcona	288	259
Crawford	137	154
Keweenaw	5	70
Luce	349	301
Roscommon	234	228

here as in the cases of corn and soybean production leaves more room for redundant spatial information to come into play, which it does in a very meaningful way. Model diagnostics for the hybrid specification are quite good. The imputed values generated here, after rounding to integers, are reported in Table 12. Unconstrained estimates give a residual total for the state of 804 based upon totals, and 998 based upon densities—rather than the corresponding value, reported previously, of 1012. Once again, employing auxiliary information supplied by totals dampens the over- and under-estimations of suppressed values generated by a model like equation (6). In terms of percentage deviation from the post-stratified county values, the most problematic cases are the two counties with the lowest head counts. Once again the estimated values seem to have little detectable impact upon the pure SAR results.

Head of Milk Cows

Analysis of milk cows data reveals that head counts by county better conform to a normal frequency distribution when $\delta = 9$ and $\gamma = \frac{1}{4}$ for totals, and $\delta = 0$ and $\gamma = \frac{1}{3}$ for densities: the Shapiro-Wilk (S-W) statistic improves

from 0.77322††† to 0.96872 for total head, and from 0.80213††† to 0.96997 for density of head, suggesting adequate conformity with a normal distribution. The accompanying MCs measuring spatial autocorrelation are: 0.38034††† and 0.50140†††, respectively, for the raw totals and densities data, and 0.41235††† and 0.51488†††, respectively, for the power-transformed totals and densities data. In other words, the frequently found moderate level of positive spatial autocorrelation is present in these power-transformed data, too. Five counties have no milk cows; but, the indicator variable differentiating these counties from those with milk cows does not have a significant regression coefficient. However, a masked outlier (Gladwin) appears to emerge through the analysis, indicating the adoption of a trivariate regression equation specification.

Restricting attention to the power-transformed data, and setting $\theta = 0$ and $\eta = \frac{1}{3}$ (rather than 2 and 1/6, the respective values needed if totals data are employed), estimation of equation (2) yields

$a = -0.01703$, $b_x = 4.43046$,
$b_{Gladwin} = -0.08735$, $R^2 = 0.917$,
S-W = 0.94348†††, MC = 0.13255††.

In other words, the geographic distribution of density of milk farms accounts for nearly 92% of the variance in the geographic distribution of head of milk cows, as well as a sizeable amount of the spatial autocorrelation displayed by this geographic distribution. Considerable strength is available to borrow from if model-based estimation for small geographic areas is to be undertaken. Approximation of the MC provided by setting $(Y_m - a - b_x X_m)$ to 0 yields

$a = -0.00935$, $b_x = 4.32160$,
$b_{Gladwin} = -0.08876$, $R^2 = 0.900$,
S-W = 0.90097†††, MC = 0.14365††,

which does not differ very much from the complete data results, once more allowing a user to make reasonable decisions about which suppressed data equation to implement.

Meanwhile, estimating a pure spatial autoregressive model, corresponding to equation (4) but for the complete data set, yields

totals: $a = 33.76262$, $\hat{\rho} = 0.69616$,
pseudo-$R^2 = 0.511$, S-W = 0.97958,
MC = 0.04232,

and

densities: $a = 0.17066$, $\hat{\rho} = 0.70967$,
pseudo-$R^2 = 0.536$, S-W = 0.98153,
MC = -0.02959.

These results reveal that moderate positive spatial autocorrelation is present in these head of milk cows data, correlation that is essentially completely accounted for by an SAR model specification, that nearby values account for roughly half of the variation in Y, and that the SAR residuals conform to a normal frequency distribution. In other words, as with corn, soybeans, and beef cattle, considerable redundant information is available to borrow from if model-based estimation for small geographic areas is to be undertaken.

Because district totals are not considered to be as reliable for milk cows, CALJACK calibration was to the state total. Accordingly, the suppressed county values estimated here (coming from six of the nine districts) are constrained to this single state total. Estimating equations (2) and (4) as well as the hybrid specification, corresponding to equation (6), in the presence of 24 suppressed data values,

Table 13: Diagnostic Statistics for Suppressed Data Estimates, Head of Milk Cows

Estimator	Mean squared prediction error (MSPE)	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
totals			
suppressed values weighted means (i.e., $n_{m,}T_k/n_{m,k}$)	62,810	0.96367†	*
equation (2)	65,542	0.90605†††	0.02310
equation (4)	310,736	0.95793††	0.16755
equation (6) [$\hat{\rho}$ = 0.27779]	63,367	0.89898†††	0.02191
densities			
suppressed values weighted means (i.e., $n_{m,}T_k/n_{m,k}$)	62,810	0.96806†	*
equation (2)	68,062	0.90067†††	0.03928
equation (4)	381,035	0.95608††	0.22986
equation (6) [$\hat{\rho}$ = 0.26531]	73,235	0.89804†††	0.03725
Note: * denotes RSSE cannot be calculated.			

which account for just 4.02% of the total number of milk cows, yields suppressed data estimates with the properties tabulated in Table 13. Interestingly these results show that borrowing strength from the number of dairy farms as a covariate or nearby location values does not necessarily improve upon the weighted mean of the suppressed values as the estimator. This finding may well be attributable to the relatively large number of small area estimates that is being generated—59 observations are being used to estimate 27 parameters—coupled with little external information being introduced by a single constraint placed upon them. Again, spatial autocorrelation alone does not furnish as much redundant information as does a good covariate. In this case the hybrid model is not clearly superior to its competitors; but neither is it obviously inferior or unacceptable according to criteria other than the MSPE. The imputed values generated here, after rounding to integers, are reported in Table 14. Unconstrained estimates give a residual total for the state of 10,464 based upon totals, and 14,578 based upon densities—rather than the corresponding value, reported previously, of 12,059. This would be roughly a 13% undercount, or a 21% overcount. Again, employing auxiliary information supplied by totals dampens the over- and under-estimations of suppressed values generated by a model like equation (6). For this head of milk cows case there appears to be only a weak tendency for a county with a relatively high percentage deviation from its post-stratified value to have a low post-stratified value, and for a county with a relatively low percentage deviation from its post-stratified value to have a high post-stratified value. As before, the impact of so many estimated values seems to have little detectable impact upon the pure SAR results.

Implications of the Michigan Data Analysis

Generally speaking there is little difference in performance between suppressed values estimation based upon totals and such estimation based upon densities in the case of the geographic distribution of specific agricultural production across Michigan, by county. But, conceptually speaking, size of county should be controlled for—larger counties have more land area available for farming. Ignoring any size effects in georeferenced data often becomes a source of variance heterogeneity. Consequently, all of the ensuing estimations are based upon densities. Parallel estimates

Table 14: Suppressed Data Estimates for Michigan Counties, Head of Milk Cows

County	Estimate based on totals	Estimate based on densities
Alger	359	411
Antrim	762	725
Baraga	381	398
Benzie	117	164
Charlevoix	841	824
Cheboygan	549	455
Dickinson	474	519
Emmet	835	796
Grand Traverse	777	716
Houghton	706	726
Iron	190	205
Kalkaska	298	353
Lake	562	568
Leelanau	828	850
Mackinac	420	465
Manistee	330	344
Marquette	271	203
Midland	849	853
Montmorency	624	614
Oakland	462	423
Otesgo	160	196
Schoolcraft	139	159
Wayne	227	254
Wexford	897	835

based upon totals have been retained for the Michigan data for comparative purposes. Adoption of this approach, which is a judgement call, avoids more anomalies and at times involves simpler power transformations. Nevertheless, useful small geographic area estimates are furnished while visual inspection of Figure 2 suggests that sufficient noise appears to be preserved in the rendered model-based estimates to continue to ensure confidentiality.

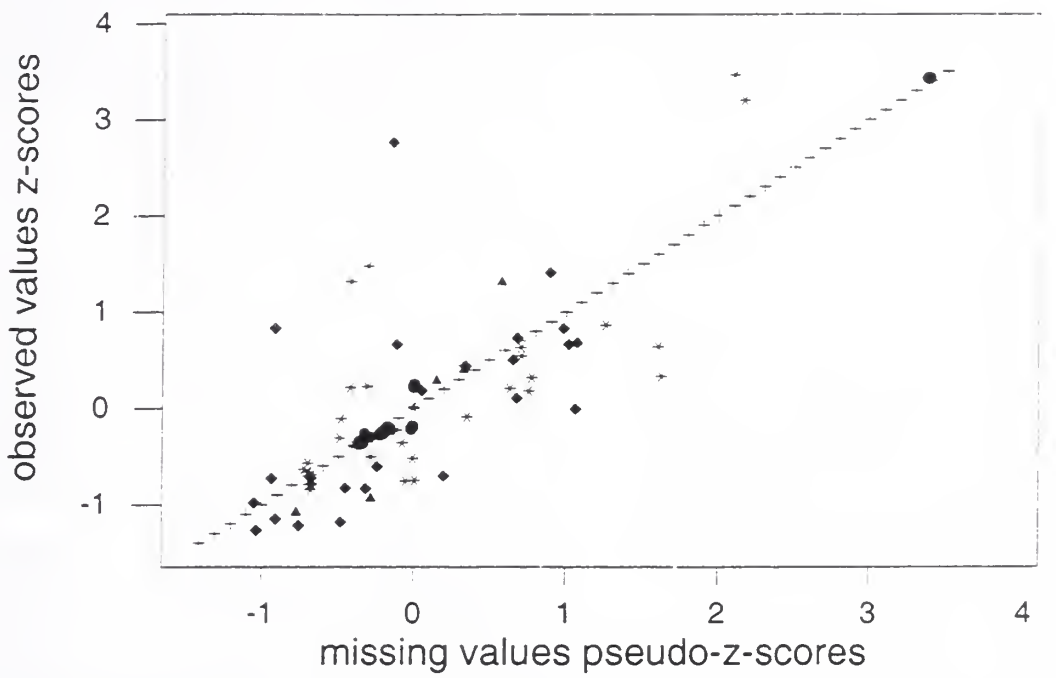
The quality of the suppressed data estimates produced by equation (6) merits additional commentary here. The scatter of points portrayed in Figure 2 reveals a moderate tendency ($r = 0.706$) for suppressed data to be predictable from the correlation structures latent in post-stratification by county of the principal crops data. Within this context, the smallest values tend to be the most difficult ones to estimate, whereas the largest values tend to be the easiest ones to estimate. The following rank correlation coefficients corroborate this contention. They have been calculated for the rank size of observed values versus the rank absolute percent deviation estimates are from these values, using the densities data:

acres of corn:	$r = -0.76$
acres of soybeans:	$r = -0.56$
bushels of corn:	$r = -0.47$
bushels of soybeans:	$r = -0.24$
head of beef:	$r = -0.70$
head of milk cows:	$r = -0.50$

Of note is that for this small geographic area estimation work, the best contribution made by spatial autocorrelation occurs when the employed covariate supplies only a moderate amount of redundant information.

Several conspicuous features of the findings

Figure 2: Bivariate plot of the observed versus the SAR-imputed values for Michigan. Circles denote corn production (acres or bushels), asterisks denote soybean production (acres or bushels), diamonds denote head of beef cattle, and triangles denote head of milk cows.



are noteworthy here, too. First, the hybrid model—equation (6)—performs best overall. Second, the number of farms of a particular type furnishes a very good covariate for inclusion in equation (6). Third, even modest amounts of geographic smoothing tend to improve estimates. Fourth, more subtotal constraints on the estimates seem to allow more reasonably good small area estimates in the presence of relatively few observations. And fifth, suppressing all figures for which the number of farms is less than 15 appears to satisfy the two operational confidentiality rules of suppression: (1) when there are c_{rule} or fewer farms of a specific type, where c_{rule} is established by the confidentiality algorithm; and, (2) when the difference between total production minus production of the second largest farm is less than (1+p) times the production of the largest farm (the p-percent protection rule). A tabulation of counties

Table 15: Numbers of Counties Whose CALJACK Farm Counts and Production Distributions Satisfy the Various Suppression Rules

commodity	<15 farms	<15 & $\leq c_{rule}$ farms & not p-% rule	<15 farms & p-% rule & not $\leq c_{rule}$	<15 & $\leq c_{rule}$ farms & p-% rule
acres of corn	13	2	0	5
bushels of corn	13	1	0	6
acres of soybeans	20	2	1	9
bushels of soybeans	20	3	1	8
head of beef cattle	5	0	0	0
head of milk cows	24	0	1	1

satisfying these three rules appears in Table 15. Therefore, not only does the criterion of employing small area estimation when the number of farms is less than 15 strengthen the resulting statistical estimates, but it also helps ensure that confidentiality is preserved.

The preceding analyses also suggest several recommendations. First, and foremost, a useful guideline appears to be: when a complementary county is needed, choose the one having the smallest commodity value and the largest estimated sampling variance. Second, estimation of the nature and degree of spatial autocorrelation present using residuals from a regression based solely upon the unsuppressed data, supplemented with zeroes for all suppressed data, appears to be quite reliable. The robustness of both of these conjectures needs further research.

Three additional themes in need of subsequent research are: (1) whether or not the minimum value for each small area estimate should be set to 1, as is done here in equations (1)-(6); (2) bootstrap and jackknife variance estimates for the small area estimator furnished by equation (6); (3) trade-offs between efficiency/precision gains and MSPE; and, (4) whether or not an indicator variable always should be included to differentiate between counties with and without production.

THE 1997 TENNESSEE DATA ANALYSIS

The major agricultural commodities in Tennessee are corn and soybeans, both measured in acres and in bushels, and cotton, measured in acres and in bales. The state of Tennessee is partitioned into n=95 counties, which are regionalized into K=6 districts. Data were collected for 66,022 individual farm operations (which translate to 79,995 CALJACK-equivalent farms) and recorded for the county of each farm's headquarters. NASS district totals for only these three crops are considered reliable enough to be used in calibration—these are the figures to which CALJACK calibrated census results. The NASS state totals for all other commodities were considered sufficiently reliable to be used as calibration benchmarks. Post-stratification by county reveals the tabulations reported in Table 16. The totals for the suppressed values (i.e., figures for counties having $0 < n_{f,i} \leq 14$) appear in Table 17. One needs to recall that these are the totals, T_k, to which suppressed county commodity estimates must sum by district.

Table 16: Post-stratification of Tennessee Farm Types by County, and Affiliated County Suppression Statistics

Agricultural commodity	Corn	Soybeans	Cotton
number of farms	8,721	6,401	1,343
number of counties with 0 farms	0	12	69
% of commodity suppressed	acres: 0.54 bushels: 0.40	acres: 1.57 bushels: 1.36	acres: 2.86 bushels: 2.34
number of counties with 1-14 farms	District 30: 2 District 40: 2 District 50: 3 District 60: 3	District 30: 4 District 40: 8 District 50: 8 District 60: 14	District 10: 2 District 20: 5 Districts 30 & 40: 5 District 50: 2

Of note is that a trivariate regression model is employed here, together with only the county densities specification, requiring the indicator variable for 0-production counties regardless of whether or not its regression coefficient is significant. Furthermore, a summary of the features of the variable transformations used appears in Table 18. Two features of these

Table 17: Districts, Combined Districts, or State Totals to Which Suppressed Values must Sum, for Tennessee

Geographic aggregate	Corn	Soybeans	Cotton
District 10			acres: 13,000‡ bales: 19,282‡
District 20			acres: 2,481 bales: 2,822
District 30	acres: 742 bushels: 54,605	acres: 1,919 bushels: 60,737	acres: 4,982 bales: 5,115
District 40	acres: 939 bushels: 79,578	acres: 6,536 bushels: 189,571	
District 50	acres: 1,226 bushels: 90,285	acres: 2,343 bushels: 70,654	acres: 3,140 bales: 3,081
District 60	acres: 445 bushels: 28,878	acres: 8,067 bushels: 233,431	

‡Includes the figure for arbitrarily selected Lake County, which serves as a complementary county here.
Note: Data for shaded cells are not relevant

Table 18: Parameters and Diagnostic Statistics for the Box-Cox and Box-Tidwell Variable Transformations

com-modity	δ	γ	Shapiro-Wilk (S-W)	farms counts for	θ	η	R^2
corn: acres	0	0	0.98150	corn	0.03	0	0.75
corn: bushels	0	0	0.98199	corn	0.02	0	0.73
soy-beans: acres	0.01	0	0.98932	soy-beans	0	0	0.39
soy-beans: bushels	0.30	0	0.98928	soy-beans	0	0	0.33
cotton: acres	0	0	0.95678	cotton	0	0	0.19
cotton: bales	0	0	0.94375	cotton	0	0	0.21

results meriting comment are the meaning of the exponent 0, and analyses involving the 0-production indicator variable. With regard to this first case, one needs to recall that an exponent of 0 refers to a logarithmic transformation. With regard to this second case, the S-W statistic has been calculated using only the non-zero values. In addition, the marginal R^2 values are reported here for soybeans and cotton; the R^2 value for the bivariate regression, which omits the 0-production indicator variable, has been subtracted from the R^2 value for the trivariate regression.

Imputation Results for Suppressed Values of Major Crops

Estimating equations (2) and (4) as well as the hybrid specification, corresponding to equation (6), in the presence of suppressed data values, which once more account for only a fraction of the total amounts of commodities, yields suppressed data estimates with the properties tabulated in Table 19. These results are in keeping with those for Michigan. One noteworthy difference is that the weighted means estimator performs comparatively better here, according to the MSPE criterion, than it does with the Michigan data. The expectation is that this relative advantage will disappear when prediction intervals are constructed.

The imputed values generated here, after rounding to integers, are reported in Table 20. As before, there is a weak-to-moderate tendency for the rank ordering of the percent deviation of the estimated suppressed values to be inversely related to the rank order of their magnitudes: -0.600 for acres of corn, -0.782 for bushels of corn, -0.440 for acres of soybeans, -0.334 for bushels of soybeans, -0.877 for acres of cotton, and -0.758 for bales of cotton.

Table 19: Diagnostic Statistics for Suppressed Corn, Soybeans, and Cotton Data Estimates, Based upon Densities

Estimator	Mean squared prediction error (MSPE)	Shapiro-Wilk statistic (S-W)	Relative sum of squared errors (RSSE)
acres of corn			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	35,716.5	0.97127	*
equation (2)	68,022.2	0.97708	0.29099
equation (4)	21,674.4	0.96910	0.25476
equation (6) [$\hat{\rho} = 0.76350$]	22,949.9	0.96388	0.12507
bushels of corn‡			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	1.9645	0.97358	*
equation (2)	4.1997	0.98013	0.14417
equation (4)	1.0211	0.96881	0.13559
equation (6) [$\hat{\rho} = 0.75604$]	1.4535	0.95925††	0.06845
acres of soybeans			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	85,489.2	0.97165	*
equation (2)	220,873.2	0.96423†	0.22189
equation (4)	115,272.2	0.98061	0.04664
equation (6) [$\hat{\rho} = 0.57136$]	95,311.6	0.97082	0.03376
bushels of soybeans‡			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	0.8624	0.97049	*
equation (2)	2.2165	0.96656	0.06410
equation (4)	1.3565	0.97230	0.01397
equation (6) [$\hat{\rho} = 0.60971$]	1.2846	0.98030	0.00958
acres of cotton			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	424,167	0.93455	*
equation (2)	4,180,767	0.97884	0.17565
equation (4)	275,702	0.90007††	0.02239
equation (6) [$\hat{\rho} = 0.17961$]	330,559	0.91381††	0.02138
bales of cotton			
suppressed values weighted means (i.e., $n_{m,i}T_k/n_{m,k}$)	725,244	0.93025†	*
equation (2)	8,274,983	0.97613	0.15458
equation (4)	562,685	0.92266†	0.02127
equation (6) [$\hat{\rho} = 0.00905$]	577,060	0.91912††	0.02122

‡The MSPEs for these crops need to be multiplied by $\times 10^8$.
Note: * denotes RSSE cannot be calculated.

As with the Michigan data, useful small geographic area estimates for Tennessee are furnished while sufficient noise is preserved in the rendered small area estimates to continue to ensure confidentiality (see Figure 3). There is nothing in the pattern portrayed in Figure 3 to indicate that the somewhat simple weighted averages are superior to their spatial autoregressive counterparts. Rather, two conspicuous weighted average outliers may suggest just the opposite.

CONCLUDING REMARKS

The research findings summarized in this report confirm that useful small geographic

Table 20: Suppressed Data Estimates for Tennessee Counties, Corn, Soybeans, and Cotton

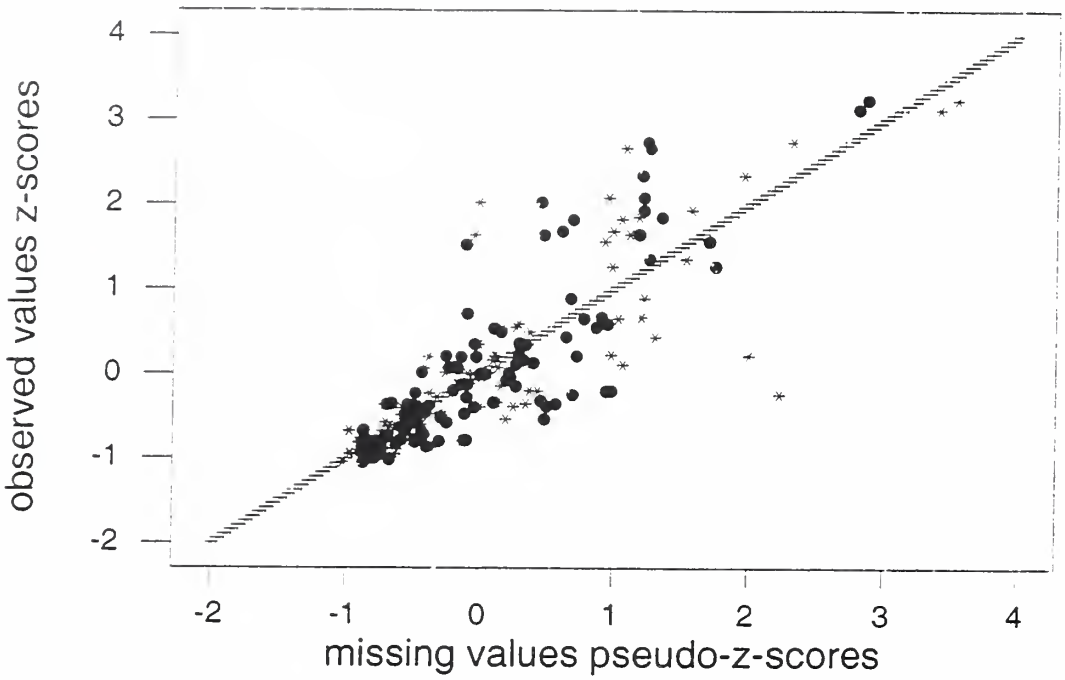
County	Corn		Soybeans		Cotton	
	acres	bush-els	acres	bush-els	acres	bales
Anderson	87	5,833				
Bradley			182	5,126		
Clay			748	22,955		
Cocke			499	14,710		
Coffee					493	435
Cumberland			113	3,300		
Davidson	224	21,786	199	5,926		
Dickson			562	17,716		
Fentress			189	5,645		
Franklin					2,647	2,646
Giles					155	136
Grainger			46	1,236		
Hamblen			1,067	32,730		
Hamilton			389	10,541		
Hardin					300	341
Hawkins			289	8,027		
Henderson					1,265	1,508
Henry					317	322
Houston	466	34,583	83	2,585		
Jackson			48	1,414		
Jefferson			1,353	40,542		
Knox			656	18,011		
Lake					10,656	16,042
Lawrence					551	530
Lewis	276	20,022	182	5,628		
Lincoln					2,108	2,146
Loudon			712	18,684		
Mcminn			1,117	31,026		
Mcnairy					458	511
Marshall			1,759	46,467		
Maury					155	136
Meigs	248	14,912	405	12,037		
Moore	715	57,792	587	15,488		
Morgan			197	5,901		
Obion					2,344	3,240
Overton			340	10,328		
Pickett	77	5,527	122	3,638		
Putnam			264	8,011		
Rhea			1,190	36,451		
Rutherford					2,014	2,168
Sequatchie	695	52,600	611	18,367		
Sevier			140	3,803		
Smith			529	16,037		
Stewart			1,092	34,807		
Trousdale			1,597	50,629		
Unicoi	120	8,133				
Van Buren	453	32,159				
Washington			22	509		
Weakley					140	140
White			506	15,465		
Wilson			1,069	30,654		

Note: Shaded cells indicate that figures have not been suppressed.

area estimates are obtainable that preserve confidentiality, allow model-based imputations to be reported, and should prove informative to the NASS debate concerning a "one-number" census utilizing numerical calibrations like CALJACK adjustments[12]. Confi-

[12]This imputation methodology is sufficiently flexible that it also can be applied to the nonresponse-CALJACK-adjusted published 1997 Census of Agricultural figures. Consider the post-stratified counts for head of beef cattle in Michigan by county. Figures for the counties of Benzie and Schoolcraft are suppressed in Table 1 of volume AC97-A-22; their respective published numbers of beef cattle farms are 28 and 17. Subtracting the published head of cattle counts for the remaining 81 counties from the published total for the state yields 862 head. The

Figure 3: Bivariate plot of the observed versus the SAR- and the weighted means-imputed values for Tennessee. Circles denote SAR-imputed values for corn production (acres or bushels), soybean production (acres or bushels), and cotton production (acres or bales). Asterisks denote weighted means-imputed values for corn production (acres or bushels), soybean production (acres or bushels), and cotton production (acres or bales).



dentiality is preserved in the sense that the actual figures for given individual farm operations are not released. Rather, only synthetic versions of information gleaned from published farming operations would be released. Presently anyone with a working knowledge of equation (6) would be able to calculate these values using current public domain data sets. If a synthetic number is too realistic, the Agency could corrupt it by infusing it with additional noise in order to mask the actual figure. Serendipitously, release of such figures by the Agency would discourage researchers from trying to generate values with equation (6). What remains to be established, then, is the minimum level of noise that synthetic numbers need to contain in order to guarantee confidentiality.

Suppression of values for the Michigan data are based upon census-adjusted number of farms, whereas suppression of values for Tennessee are based upon CALJACK-adjusted number of farms. Some counties that satisfy the cut-off value of 14 farms when the count is based strictly upon survey results fail to exceed this cut-off value after also being CALJACK-adjusted. Similarly, some counties whose number of farms is less than 14 according to the survey results exceed 14 after being CALJACK-adjusted. The number of counties that flips in each case is quite small; the number observed during this research is 4 out of the 4×83 possible Michigan county cases, and 6 out of the 3×95 possible Tennessee county cases. Nevertheless, there is a need to decide upon a sound threshold number for suppression, and to determine if this number

imputation methodology renders estimates for these two counties whose total sums to 862. They are, based upon a density analysis and constraining the minimum to be 1: 465 for Benzie County, and 397 for Schoolcraft County.

should be survey-based or CALJACK-adjusted-based, or perhaps both. A second issue meriting subsequent examination is whether or not the CALJACK optimization should begin with ones or the weights derived in conjunction with the agricultural census. The Tennessee results used in this investigation utilize the former initial weights, while the Michigan results used in this investigation utilize the latter initial weights. Of note is that the imputation methodology evaluated here is sufficiently flexible that it can be adapted to either situation.

For the most part, the study generating this report raises a number of important issues for subsequent research. Foremost is the need to establish prediction intervals for the model-based small geographic area estimates. Jackknife, bootstrap, and Monte Carlo simulation experiments are called for. Second is the need to analyze criteria for selecting a good estimator. Candidates included in this work are mean squared prediction error (MSPE), relative sum of squared errors (RSSE), and model diagnostics. Concerns here include whether or not the MSPE is biased, especially in the presence of regional constraints, and statistical theory impacts when the portion of the EM algorithm likelihood function involving the imputations cannot converge to 0. Third is the need to re-examine the notion of confidentiality when a single crop-farm is present in a county. Related to this issue is the need to guarantee a minimum level of noise in the imputed data values in order to guarantee confidentiality. Fourth is the need to automate the imputation methodology, perhaps with SAS macros.

The methodology also is accompanied by a number of caveats. For example, model-based small geographic area estimates inserted into released tables should be identified as such. A data user should be advised that because these are synthetic values, they will differ from the suppressed values by some amount. Another caution concerns the degree to which such synthetic values need to be reviewed for disclosure avoidance. Regardless of how prudently such values are handled, the fact remains that currently a knowledgeable developer and user of equations like (6) already is capable of generating the types of model-based small geographic area estimates discussed in this report. Therefore, the Agency should either take steps to release its preferred versions of such synthetic numbers, or at the very least scrutinize such numbers as part of its confidentiality assessments.

RECOMMENDATIONS

The author recommends that further research be conducted in the following topic areas to more fully assess the utility of the model-based small geographic area estimation procedure presented in this report:

> (1) establish prediction intervals for the small geographic area estimates.

A research project focusing on this theme also should address the issue of preserving a minimum amount of noise in model-based imputations in order to ensure confidentiality.

> (2) evaluate the imputation methodology for many, if not all, of the 50 states, as well as its extension to non-geographic small area estimation problems.

Completion of this second project should provide a database for identification of trade-

offs between estimator selection criteria (e.g., MSPE, RSSE, model diagnostics), which would be a third project.

> (3) identify a battery of selection criteria for discriminating between model-based small geographic area estimators.

Closely related to this third project would be

> (4) assess statistical properties of the selected estimator, such as robustness, bias, efficiency, consistency, and suppression criteria (e.g., the p-% rule), as well as statistical properties of the selection criteria identified in (3).

This project would be more mathematical statistics theoretic in nature. Another worthwhile research project would be

> (5) automation of the final version of the methodology, perhaps using SAS macros.

These five projects could be sequenced in various ways. Presumably highest priority should be given to the first one.

REFERENCES

Citro, C. 1998. "Model-based small-area estimates: the next major advance for the federal statistical system for the 21st century," *Chance*, 11: 40-41, 50.

Deville, J-C., and C-E Särndal. 1992. "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, 87: 376-382.

Griffith, D., J. Paelinck, and R. van Gastel. 1998. "The Box-Cox transformation: computational and interpretation features of the parameters," in D. Griffith and C. Amrhein, *Advances in Spatial Modelling and Methodology: Essays in Honor of Jean Paelinck*, Dordrecht: Kluwer, pp. 45-56.

Navidi, W. 1997. "A graphical illustration of the EM algorithm," *The American Statistician*, 51: 29-31.

Pannekoek, J., and T. de Waal. 1998. "Synthetic and combined estimators in statistical disclosure control," *J. of Official Statistics*, 14: 399-410.

Perry, C., R. Chhikara, F. Spears, S. Cowles, and W. Iwig. 1997. "An evaluation of list-only, reweighted, and other estimators for U.S. agricultural labor surveys," *Research Division Research Report* No. RD-97-06. Washington, D.C.: National Agricultural Statistics Service, USDA.

Stasny, E., P. Goel, C. Cooley, and L. Bohn. 1995. "Modeling county-level crop yield with spatial correlations among neighboring counties," *Technical Report* No. 570. Colombus, Ohio: Department of Statistics, The Ohio State University.

USDA. 1999. *Understanding USDA Crop Forecasts*. Washington, D.C.: Miscellaneous Publication No. 1554, National Agricultural Statistics Service & Office of the Chief Economist, World Agricultural Outlook Board.+

APPENDIX A. Sample SAS code: The Michigan Beef Cattle Case

```
*************************************************************
*                                                           *
* INPUT FILES - COUNTY-REFERENCED AGRICULTURAL DATA          *
*             - EIGENVALUES OF GEOGRAPHIC WEIGHTS MATRIX FOR *
*               ALL COUNTIES                                *
*             - EIGENVALUES OF GEOGRAPHIC WEIGHTS MATRIX FOR *
*               COUNTIES WITH SUPPRESSED VALUES             *
*             - GEOGRAPHIC WEIGHTS MATRIX FOR ALL COUNTIES  *
*             - OUTPUT FILE CONTAINING VALUES BEING         *
*               SUPPRESSED AND THEIR SPATIAL STATISTICAL    *
*               MODEL ESTIMATES                             *
*                                                           *
*************************************************************;

FILENAME NASS  'O:\CALJACK\MICH-CALJACK.DAT';
FILENAME EIGEN 'O:\CALJACK\MICHIGAN CENSUS DATA\MICHIGAN.EIG';
FILENAME EIGENMIS 'O:\CALJACK\MICH-MISS-BEEF.EIG';
FILENAME CONN 'O:\CALJACK\MICHIGAN CENSUS DATA\MICHIGAN.CON';
FILENAME OUTFILE 'C:\SAS-BEEF.OUT';
OPTIONS LINESIZE=72;

**************
*            *
* INPUT DATA *
*            *
**************;

DATA STEP0;
     INFILE NASS;
     INPUT COUNTY DISTRICT ACRES NO WN
               NC WNFCORN CORNA CORNB NS WNFSOY SOYA SOYB
               NBF WNFBEEF BEEF NMLK WNFMILK MILK;
CROP= BEEF; N=WNFBEEF;
DENOM = ACRES;
AY=  0 ; BY=2;
AX=  0 ; BX=2;
Y = (CROP/DENOM + AY)**(1/BY);
X = (N/DENOM + AX)**(1/BX);
YHOLD=CROP;
RUN;
PROC SORT OUT=STEP0(REPLACE=YES); BY COUNTY; RUN;

*********************
*                   *
* INPUT EIGENVALUES *
*                   *
*********************;

DATA STEP1;
      INFILE EIGEN;
      INPUT NUM EIGEN;
RUN;

***********************************************************
*                                                         *
* IDENTIFICATION OF COUNTIES WHOSE VALUES ARE SUPPRESSED  *
* AND REPLACEMENT OF THESE SUPPRESSED VALUES WITH 0       *
*                                                         *
***********************************************************;

DATA STEP1(REPLACE=YES);
     SET STEP1 (KEEP=EIGEN);
     SET STEP0 (KEEP=COUNTY DISTRICT Y X YHOLD N AY BY DENOM);
LAMBDA=EIGEN;
```

```
IF COUNTY=  3 THEN IM3  =1; ELSE IM3  =0;
IF COUNTY= 39 THEN IM39 =1; ELSE IM39 =0;
IF COUNTY= 83 THEN IM83 =1; ELSE IM83 =0;
IF COUNTY= 95 THEN IM95 =1; ELSE IM95 =0;
IF COUNTY=143 THEN IM143=1; ELSE IM143=0;

IMISS = IM3 + IM39 + IM83 + IM95 + IM143;
IF IMISS=1 THEN Y=0;
RUN;
DATA TEMPN;
     SET STEP1;
IF IMISS=0 THEN DELETE;
RUN;
PROC PRINT; VAR COUNTY DISTRICT N; RUN;
PROC MEANS SUM; VAR N; RUN;

**********************************
*                                *
* INPUT GEOGRAPHIC WEIGHTS MATRIX *
*                                *
**********************************;

DATA STEP1 (REPLACE=YES);
     SET STEP1;
     INFILE CONN;
     INPUT NAME C1-C83;
     ARRAY CONY{83} CY1-CY83;
     ARRAY CONX{83} CX1-CX83;
     ARRAY CONI0{83} C0I1-C0I83;
     ARRAY CON{83} C1-C83;
     CSUM = 0;
     DO I=1 TO 83;
          CSUM = CSUM + CON{I};
          CONY{I} = Y*CON{I};
          CONX{I} = X*CON{I};
     END;
RUN;

************************************************************
*                                                          *
* CONSTRUCTION OF SPATIAL LAG VARIABLES FROM CONY, CONX    *
* AND CONI0, WHEN AN INDICATOR VARIABLE FOR 0 PRODUCTION   *
* IS NEEDED                                                *
*                                                          *
************************************************************;

PROC MEANS DATA=STEP1 NOPRINT;
     VAR CY1-CY83 CX1-CX83;
     OUTPUT OUT=CYOUT1 SUM=CY1-CY83 CX1-CX83;
RUN;
PROC TRANSPOSE DATA=CYOUT1 PREFIX=CY OUT=CYOUT2Y;
     VAR CY1-CY83;
RUN;
PROC TRANSPOSE DATA=CYOUT1 PREFIX=CX OUT=CYOUT2X;
     VAR CX1-CX83;
RUN;
PROC TRANSPOSE DATA=STEP1 PREFIX=TLAM OUT=CYOUT3;
     VAR LAMBDA;
RUN;
DATA STEP2(REPLACE=YES);
     INFILE EIGENMIS;
     INPUT LAMBDAWM;
RUN;
PROC TRANSPOSE DATA=STEP2 PREFIX=TLAMM OUT=CYOUT4;
     VAR LAMBDAWM;
RUN;

DATA STEP1 (REPLACE=YES);
```

```
     SET STEP1;
     IF _N_=1 THEN SET CYOUT3;
     IF _N_=1 THEN SET CYOUT4;
     SET CYOUT2Y;
     SET CYOUT2X;
WY = CY1/CSUM;
WX = CX1/CSUM;
RUN;

************************************************
*                                              *
* BIVARIATE REGRESSION USING UNSUPPRESSED VALUES *
*                                              *
************************************************;

DATA TEMP0A;
     SET STEP1;
IF IMISS=1 THEN Y='.';
RUN;
PROC REG; MODEL Y = X; OUTPUT OUT=TEMP0B P=YHAT; RUN;
DATA TEMP0B(REPLACE=YES);
     SET TEMP0B;
IF IMISS=0 THEN DELETE;
RUN;

PROC SORT DATA=TEMP0B OUT=TEMP0B(REPLACE=YES); BY COUNTY; RUN;
PROC PRINT; VAR COUNTY DISTRICT YHAT; RUN;
DATA LABELS;
     SET TEMP0B(KEEP=DISTRICT COUNTY);
RUN;

**************
*            *
* YM = YMBAR *
*            *
**************;
DATA STEP3;
     SET STEP1;
YMBAR=Y;
IF IMISS=1 THEN YMBAR=(1012*N/47/DENOM - AY)**(1/BY);
X0=1;
RUN;
PROC REG; MODEL YMBAR=X0/NOINT; OUTPUT OUT=STEP0B R=YRESID; RUN;
PROC UNIVARIATE NORMAL; VAR YRESID; RUN;
DATA STEP3A;
     SET STEP3;
YMBAR = DENOM*(YMBAR**BY + AY);
IF IMISS=0 THEN DELETE;
RUN;
PROC PRINT; VAR DISTRICT COUNTY Y YMBAR; RUN;

*********************************
*                               *
* THE EM-TYPE STATISTICAL MODEL *
*                               *
*********************************;

PROC NLIN DATA=STEP1 MAXITER=500 METHOD=MARQUARDT;
     PARMS A=-1.2 BX=3.0
           M3=1 M39=1 M83=1 M95=1;
     BOUNDS 0<M3, 0<M39, 0<M83, 0<M95;
T=1012;
     MODEL Y = A + BX*X
                 - ( IM3  *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     IM39 *((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     IM83 *((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     IM95 *((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     IM143*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) );
```

```
     OUTPUT OUT=TEMP5 PRED=YHAT R=YRESID
                      PARMS=A BX MISS1-MISS4;
RUN;
PROC UNIVARIATE DATA=TEMP5 NORMAL; VAR YRESID; RUN;
DATA TEMP5(REPLACE=YES);
     SET TEMP5;
IF IMISS=0 THEN DELETE;
RUN;
PROC PRINT; VAR DISTRICT COUNTY Y YHAT; RUN;
DATA TEMP6;
    SET TEMP5(KEEP = MISS1-MISS4);
MISS5 = 1;
IF _N_>1 THEN DELETE;
ARRAY MISSING1{5} MISS1-MISS5;
     SUM = 0;
     DO I=1 TO 5;
        SUM = SUM + MISSING1{I};
     END;
     DO I=1 TO 5;
        MISSING1{I} = 1 + (1012-5)*MISSING1{I}/SUM;
     END;
RUN;
PROC TRANSPOSE OUT=TEMP7 PREFIX=MISSEM; VAR MISS1-MISS5; RUN;
DATA TEMP7(REPLACE=YES);
     SET TEMP7;
     SET LABELS;
RUN;
PROC SORT OUT=TEMP7(REPLACE=YES); BY COUNTY; RUN;

*************************************************************
*                                                           *
* THE PURE SAR SPATIAL STATISTICAL MODEL USING MATRIX W     *
*                                                           *
*************************************************************;
PROC NLIN DATA=STEP1 MAXITER=500 METHOD=MARQUARDT;
     PARMS RHO=0.9 A=35
           M3=1 M39=1 M83=1 M95=1;
     BOUNDS -1.31498<RHO<0.999999, 0<M3, 0<M39, 0<M83, 0<M95;
T=1012;
     ARRAY LAMBDAJ{83} TLAM1-TLAM83;
     ARRAY LAMBDAMJ{5} TLAMM1-TLAMM5;
     JACOB = 0;
     DERJ = 0;
     DO I=1 TO 83;
          JACOB = JACOB + LOG(1 - RHO*LAMBDAJ{I});
          DERJ = DERJ + -LAMBDAJ{I}/(1 - RHO*LAMBDAJ{I});
     END;
     JMISS=0;
     DERJM = 0;
     DO I=1 TO 5;
          JMISS = JMISS + LOG(1 - RHO*LAMBDAMJ{I});
          DERJM = DERJM + -LAMBDAMJ{I}/(1 - RHO*LAMBDAMJ{I});
     END;

     J=EXP((JACOB-JMISS)/(83-5));
     DERJ = -(DERJ - DERJM)/(83-5);

     ZY = Y/J;

     MODEL ZY = (RHO*(CY1
                 + ( C2 *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C20*((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C42*((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C48*((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C72*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                        )/CSUM

           + A*(1 - RHO)
```

```
                - ( IM3  *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM39 *((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM83 *((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM95 *((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM143*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                       )/J;
     OUTPUT OUT=TEMP8 PRED=YHAT R=YRESID PARMS=RHO A MISS1-MISS4;

     DER.RHO = ((RHO*(CY1
                + ( C2 *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C20*((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C42*((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C48*((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C72*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                       )/CSUM

          + A*(1 - RHO)
                - ( IM3  *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM39 *((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM83 *((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM95 *((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM143*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )

          - Y)*DERJ + (CY1

                + ( C2 *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C20*((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C42*((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C48*((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C72*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                       )/CSUM
          - A)/J;
RUN;
PROC UNIVARIATE DATA=TEMP8 NORMAL; VAR YRESID; RUN;
DATA TEMP8(REPLACE=YES);
     SET TEMP8;
IF IMISS=0 THEN DELETE;
RUN;
PROC PRINT; VAR DISTRICT COUNTY Y YHAT; RUN;
DATA TEMP9;
    SET TEMP8(KEEP = MISS1-MISS4);
MISS5 = 1;
IF _N_>1 THEN DELETE;
ARRAY MISSING1{5} MISS1-MISS5;
     SUM = 0;
     DO I=1 TO 5;
        SUM = SUM + MISSING1{I};
     END;
     DO I=1 TO 5;
        MISSING1{I} = 1 + (1012-5)*MISSING1{I}/SUM;
     END;
RUN;
PROC TRANSPOSE OUT=TEMP10 PREFIX=MISSSP; VAR MISS1-MISS5; RUN;
DATA TEMP10(REPLACE=YES);
     SET TEMP10;
     SET LABELS;
RUN;
PROC SORT OUT=TEMP10(REPLACE=YES); BY COUNTY; RUN;

*****************************************************
*                                                   *
* THE SAR SPATIAL STATISTICAL MODELS USING MATRIX W *
*                                                   *
*****************************************************;
PROC NLIN DATA=STEP1 MAXITER=500 METHOD=MARQUARDT;
     PARMS RHO=0.9 A=35 BX=3
           M3=1 M39=1 M83=1 M95=1;
     BOUNDS -1.31498<RHO<0.999999, 0<M3, 0<M39, 0<M83, 0<M95;
```

```
T=1012;
     ARRAY LAMBDAJ{83} TLAM1-TLAM83;
     ARRAY LAMBDAMJ{5} TLAMM1-TLAMM5;
     JACOB = 0;
     DERJ = 0;
     DO I=1 TO 83;
          JACOB = JACOB + LOG(1 - RHO*LAMBDAJ{I});
          DERJ = DERJ + -LAMBDAJ{I}/(1 - RHO*LAMBDAJ{I});
     END;
     JMISS=0;
     DERJM = 0;
     DO I=1 TO 5;
          JMISS = JMISS + LOG(1 - RHO*LAMBDAMJ{I});
          DERJM = DERJM + -LAMBDAMJ{I}/(1 - RHO*LAMBDAMJ{I});
     END;

     J=EXP((JACOB-JMISS)/(83-5));
     DERJ = -(DERJ - DERJM)/(83-5);

     ZY = Y/J;

     MODEL ZY = (RHO*(CY1
                 + ( C2 *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C20*((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C42*((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C48*((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C72*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                      )/CSUM

          + A*(1 - RHO) + BX*(X - RHO*WX)
                - ( IM3  *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM39 *((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM83 *((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM95 *((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM143*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                      )/J;
     OUTPUT OUT=TEMP2 PRED=YHAT R=YRESID
                      PARMS=RHO A BX MISS1-MISS4;

     DER.RHO = ((RHO*(CY1
                 + ( C2 *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C20*((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C42*((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C48*((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                     C72*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                      )/CSUM
          + A*(1 - RHO) + BX*(X - RHO*WX)
                - ( IM3  *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM39 *((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM83 *((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM95 *((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    IM143*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )

          - Y)*DERJ + (CY1

                + ( C2 *((1 + M3 *(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C20*((1 + M39*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C42*((1 + M83*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C48*((1 + M95*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) +
                    C72*((1 +   1*(T - 5)/(M3+M39+M83+M95+1))/DENOM + AY)**(1/BY) )
                                                                      )/CSUM
          - A - BX*WX)/J;
RUN;

*******************************************************
*                                                     *
* CONVENTIONAL HOMOGENEITY OF VARIANCE DIAGNOSTIC PLOT *
* FOR THE ESTIMATED SAR MODEL                         *
```

```
*                                                    *
*****************************************************;

PROC GPLOT; PLOT YRESID*YHAT; RUN;
PROC REG; MODEL Y=YHAT; RUN;
PROC UNIVARIATE DATA=TEMP2 NORMAL; VAR YRESID; RUN;
DATA TEMP2(REPLACE=YES);
     SET TEMP2;
IF IMISS=0 THEN DELETE;
RUN;
PROC PRINT; VAR DISTRICT COUNTY Y YHAT; RUN;
DATA TEMP3;
    SET TEMP2(KEEP = MISS1-MISS4);
MISS5 = 1;
IF _N_>1 THEN DELETE;
ARRAY MISSING1{5} MISS1-MISS5;
     SUM = 0;
     DO I=1 TO 5;
        SUM = SUM + MISSING1{I};
     END;
     DO I=1 TO 5;
        MISSING1{I} = 1 + (1012-5)*MISSING1{I}/SUM;
     END;
RUN;
PROC TRANSPOSE OUT=TEMP4 PREFIX=MISS; VAR MISS1-MISS5; RUN;
DATA TEMP4(REPLACE=YES);
     SET TEMP4;
     SET LABELS;
RUN;
PROC SORT OUT=TEMP4(REPLACE=YES); BY COUNTY; RUN;

DATA STEP4(REPLACE=YES);
     SET STEP1;
IF IMISS=0 THEN DELETE;
RUN;
PROC SORT DATA=STEP4 OUT=STEP4(REPLACE=YES); BY COUNTY; RUN;
DATA TEMP4(REPLACE=YES);
     SET TEMP4;
     SET STEP4(KEEP=DISTRICT COUNTY X YHOLD);
     SET STEP3A(KEEP=YMBAR);
     SET TEMP7;
     SET TEMP10;
YMBAR = ROUND(YMBAR,1);
YMISSEM = ROUND(MISSEM1,1);
YMISSSP = ROUND(MISSSP1,1);
YMISSAR=ROUND(MISS1,1);
MSEMBAR = (YHOLD - YMBAR)**2;
MSESP = (YHOLD - YMISSSP)**2;
MSEEM = (YHOLD - YMISSEM)**2;
MSEAR = (YHOLD-YMISSAR)**2;
RUN;

*******************************************************
*                                                     *
* PRINTING OF THE VARIOUS SUPPRESSED VALUES ESTIMATES *
*                                                     *
*******************************************************;

PROC PRINT; VAR COUNTY DISTRICT YHOLD YMBAR YMISSSP YMISSEM YMISSAR; RUN;

*******************************************************
*                                                     *
* FOR CHECKING THAT THE CONSTRAINED SUPPRESSED VALUES *
* ESTIMATES HAVE THE CORRECT SUM                      *
*                                                     *
*******************************************************;

PROC MEANS SUM; VAR YHOLD YMBAR YMISSSP YMISSEM YMISSAR; RUN;
```

```
PROC UNIVARIATE DATA=TEMP4; VAR MSEMBAR MSESP MSEEM MSEAR; RUN;
PROC GPLOT DATA=TEMP4; PLOT YHOLD*YMISSAR; RUN;

DATA FINAL;
     SET TEMP4;
YCOMP = ABS(1 - YMISSAR/YHOLD);
RUN;
PROC RANK OUT=FINAL(REPLACE=YES); VAR YHOLD YCOMP; RANKS RYHOLD RYCOMP; RUN;
PROC CORR; VAR RYHOLD RYCOMP; RUN;
PROC PRINT; VAR COUNTY DISTRICT YHOLD YMISSAR YCOMP RYCOMP; RUN;

*******************************
*                             *
* CREATION OF THE OUTPUT FILE *
*                             *
*******************************;

DATA _NULL_;
     SET FINAL;
     FILE OUTFILE;
     PUT COUNTY DISTRICT YHOLD YMISSAR;
RUN;

***********************************************
*                                             *
* THE UNCONSTRAINED EM-TYPE STATISTICAL MODEL *
*                                             *
***********************************************;

PROC NLIN DATA=STEP1 MAXITER=500 METHOD=MARQUARDT;
     PARMS A=-1.2 BX=3.0
           M3=202 M39=202 M83=202 M95=202 M143=202;
     BOUNDS 8<M3, 8<M39, 8<M83, 8<M95, 8<M143;

     MODEL Y = A + BX*X
     - ( IM3 *(M3/DENOM + AY)**(1/BY) + IM39 *(M39/DENOM + AY)**(1/BY) +
         IM83*(M83/DENOM + AY)**(1/BY) +
         IM95*(M95/DENOM + AY)**(1/BY) + IM143*(M143/DENOM + AY)**(1/BY) );
     OUTPUT OUT=TEMP25 PRED=YHAT R=YRESID
                       PARMS=A BX MISS1-MISS5;

RUN;
PROC UNIVARIATE DATA=TEMP25 NORMAL; VAR YRESID; RUN;
DATA TEMP25(REPLACE=YES);
     SET TEMP25;
IF IMISS=0 THEN DELETE;
RUN;
PROC PRINT; VAR DISTRICT COUNTY Y YHAT; RUN;
PROC TRANSPOSE OUT=TEMP26 PREFIX=MISSEM; VAR MISS1-MISS5; RUN;
DATA TEMP26(REPLACE=YES);
     SET TEMP26;
     SET LABELS;
MISSEM = ROUND(MISSEM1,1);
RUN;

*****************************************************
*                                                   *
* SUM OF UNCONSTRAINED SUPPRESSED VALUES ESTIMATES *
*                                                   *
*****************************************************;

PROC MEANS SUM; VAR MISSEM; RUN;
```